HAL
open science

Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression

Caroline Bernard-Michel, Sylvain Douté, Laurent Gardes, Stéphane Girard

To cite this version:

Caroline Bernard-Michel, Sylvain Douté, Laurent Gardes, Stéphane Girard. Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression. [Research Report] 2007, pp.91. inria-00187444v1

HAL Id: inria-00187444

https://inria.hal.science/inria-00187444v1

Submitted on 14 Nov 2007 (v1), last revised 15 Nov 2007 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression

Caroline Bernard-Michel — Sylvain Douté — Laurent Gardes — Stéphane Girard

N° ????

Novembre 2007

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--????--FR+ENG

Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression

Caroline Bernard-Michel*, Sylvain Douté †, Laurent Gardes*, Stéphane Girard*

Thème COG — Systèmes cognitifs
Projets MISTIS

Rapport de recherche n° ???? — Novembre 2007 — 91 pages

Abstract: Visible and near infrared imaging spectroscopy is a key remote sensing technique to study and monitor planet Mars. Indeed it allows the detection, mapping and characterization of minerals as well as volatile species that often constitute the first step toward the resolution of key climatic and geological issues. These tasks are carried out by the spectral analysis of the solar light reflected in different directions by the materials forming the top few millimeters or centimeters of the ground. The chemical composition, granularity, texture, physical state, etc. of the materials determine the morphology of the hundred thousands spectra that typically constitute an image. Radiative transfer models simulating the propagation of solar light through the Martian atmosphere and surface and then to the sensor aim at evaluating numerically the direct and quantitative link between parameters and spectra. Then techniques must be applied in order to reverse the link and evaluate the properties of atmospheric and surface materials from the spectra. Processing all the pixels of an image finally provides physical and structural maps. We use a regularized version of SIR method (K.C. Li, Sliced Inverse Regression for dimension reduction, *Journal of the American Statistical Association*, 86:316-327, 1991) combined to a linear interpolation to reverse the previous numerical link. For that purpose we first generate numerous corresponding pairs of parameters - synthetic spectra by direct radiative transfer modeling in order to constitute a learning database. The SIR step allows to reduce the dimension of the spectra (usually 184 wavelengths) in order to overcome the curse of dimensionality. Then, a linear interpolation is used to relate the reduced components of a spectrum to a given physical parameter value. Such inverted link is applied to a real dataset of hyperspectral images collected by the OMEGA instrument (Mars Express mission).

Key-words: Dimension reduction, Sliced Inverse Regression (SIR), hyperspectral images, physical modeling, regularization, Mars

* MISTIS - INRIA Rhône-Alpes
† Laboratoire de Planétologie de Grenoble

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier (France)
Téléphone : +33 4 76 61 52 00 —Télécopie +33 4 76 61 52 52

Estimation des paramètres physiques de la calotte sud de Mars à partir d'images hyperspectrales: utilisation de la méthode SIR

Résumé : La spectroscopie visible et infrarouge est une technique clé de la télédetection pour étudier la surface des planètes. Elle permet en effet la détection, la cartographie et la caractérisation des minéraux, ainsi que des espèces volatiles et constitue ainsi un premier pas pour une interprétation climatique et géologique des surfaces planétaires. Ces différentes analyses sont réalisées à partir de l'analyse spectrale de la lumière solaire refléchie par les différents matériaux présents à la surface de la planète sur une épaisseur de quelques millimètres voire quelques centimètres. En effet, la compostion chimique, la granulométrie, la texture, l'état physique, etc... des matériaux déterminent la morphologie des milliers de spectres qui constituent typiquement une image. Le modèle de transfert radiatif simule la propagation de la lumière à travers l'atmosphère et la surface martienne et permet ainsi d'évaluer analytiquement le lien direct entre les spectres et les paramètres. Diverses techniques peuvent être alors utilisées pour inverser ce lien et estimer les propriétés physiques des matériaux en surface à partir du spectre. Ainsi il est possible d'obtenir une cartographie de divers paramètres physiques de la planète Mars.
Nous proposons dans ce rapport d'utiliser une version régularisée de la regression inverse par tranches (K.C. Li, Sliced Inverse Regression *Journal of the American Statistical Association*, 86:316-327, 1991) associée à une interpolation linéaire pour estimer les propriétés physiques du sol martien. Dans un premier temps, des spectres synthétiques sont simulés par le modèle de transfert radiatif en faisant varier les différents paramètres du modèle sur une grille de valeurs représentatives de la calotte sud de Mars, ce qui constitue une base d'apprentissage de spectres. Ensuite, la méthode SIR est appliquée pour réduire la dimension des spectres (184 longueurs d'ondes) et pallier au flau de la dimension. Une technique d'interpolation linéaire est alors utilisée pour caractériser la relation fonctionnelle entre spectres réduits et paramètres. Une application est présentée pour des images hyperspectrales réelles de la planète Mars recueillies par l'instrument OMEGA de la mission Mars Express.

Mots-clés : Réduction de dimension, Regression inverse par tranches (SIR), images hyperspectrales, modélisation physique, régularisation, Mars

Contents

1 Introduction **5**
1.1 Background information . . . 5
1.1.1 Mars express mission - OMEGA . . . 5
1.1.2 Spectrometer and hyperspectral images . . . 5
1.2 The Inverse Problem . . . 6
1.3 Data . . . 8
1.3.1 Hyperspectral Images from Mars . . . 8
1.3.2 Learning databases . . . 9

2 Methodology **13**
2.1 Current approach: Nearest neighbors algorithm . . . 13
2.2 Proposed approach: Sliced Inverse Regression . . . 16
2.2.1 Sliced Inverse Regression . . . 16
2.2.2 Regularized Sliced Inverse Regression . . . 21
2.2.3 Estimation of the functional relationship . . . 23

3 Validation on simulations **25**
3.1 Simulation of a test data . . . 25
3.2 Validation criteria . . . 26
3.2.1 Sliced Inverse Regression criterion . . . 27
3.2.2 Normalized RMSE criterion . . . 27
3.3 Choice of the regularization parameter . . . 28
3.4 Influence of the slices . . . 32
3.5 Comparisons between methods . . . 32
3.6 Choice of the learning database . . . 35
3.7 How to deal with dependent parameters? . . . 40
3.8 Final results . . . 41

4 Application to real data **43**
4.1 How to validate results for a real data? . . . 43
4.2 Masking some of the wavelengths? . . . 44
4.3 Selection of the learning database . . . 46

4.4 Final GRSIR methodology 47
4.5 Results 48

5 Conclusion **69**

A Selected Wavelengths **71**

B Principal component analysis (PCA) **73**

C Functional relationship **77**

D SIR weights **81**

E Choice of the regularization parameter **83**

Chapter 1

Introduction

1.1 Background information

1.1.1 Mars express mission - OMEGA

In June 2003, the first European space craft, Mars express, has been sent in orbit around Mars by the European Space Agency (ESA). It carries seven orbiter instruments (surface and sub surface instruments, atmosphere and ionosphere instruments, radio ...) and one lander, Beagle 2, that has unfortunately been officially declared lost in 2004. From a general point of view, the aims of Mars express mission are to image the entire surface of Mars at high resolution, to produce maps of the mineral composition of the surface but also of the composition of the atmosphere and of the polar caps, to determine the structure of sub-surface and to understand the effect of the atmosphere on the surface. In the future, the analysis of these results should help to understand the geological and climatological history of Mars but also of our own planet, the earth.
In this report, we will concentrate on datasets collected by the French spectro-imaging instrument: OMEGA (observatoire pour la minéralogie, l'eau, la glace et l'activité). OMEGA has been developed by IAS and LESIA (Observatoire de Paris) with the support of CNES, and a participation of IFSI (Italy) and IKI (Russia). This visible and infrared mineralogical mapping spectrometer should observe most of the Martian surface. It records the visible and infrared light reflected from the planet's surface in the wavelength range 0.5-5.2 microns and with a ground resolution varying from 350 m to 10 km. After physical, statistical and computational treatment, such observations allow the characterization and mapping of the main minerals of Mars surface and also measures aspects of atmospheric composition. Thus the Martian geology and its evolution should be better understood.

1.1.2 Spectrometer and hyperspectral images

To summarize briefly, a spectrometer is an instrument that measures the amount of light reflected or emitted by a surface for different wavelengths. The curve of reflectance as a function of wavelength is called the light spectrum or also the spectral signature of the observed surface. OMEGA spectrometer collects the spectrum for each pixel of the observed

surface and for more than 300 wavelengths. In this report, we will only consider the 184 wavelengths given in appendix A. The data generated for a given portion of planetary surface is then a data cube, called hyperspectral image. It is composed of two spatial dimensions, representing the studied surface, and one spectral dimension composed of the wavelengths (see figure 1.1.1). As a first step, hyperspectral images are generally used for detection of compounds, classification and segmentation in order to define spectrally homogeneous units of terrain on the planet. Then spectra are analyzed more precisely to determine the exact chemical composition, granularity, texture, physical state,... of the materials found in these areas. In this report, we will address the estimation of the physical properties of Mars surface and ices from the thousands of spectral pixels that were acquired over the south polar cap of Mars using Sliced Inverse Regression [15]. The general context of this study is well known as an inverse problem.

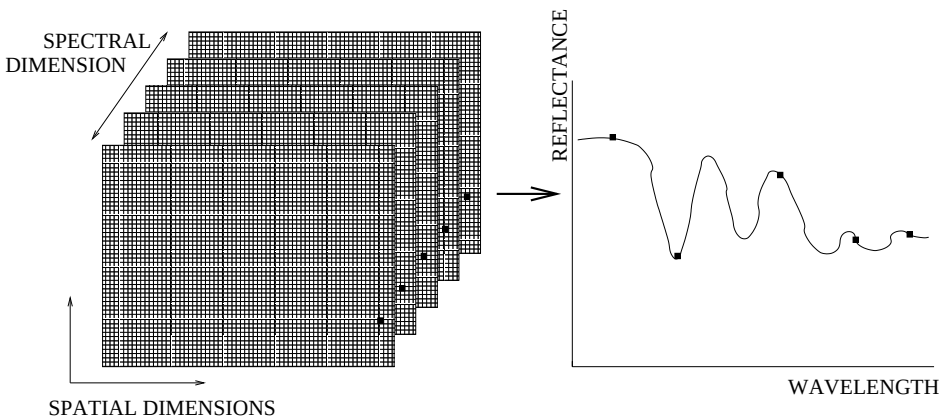


Figure 1.1.1: Hyperspectral image. Source: BRGM

1.2 The Inverse Problem

Physical properties of a surface such as the chemical composition, the granularity or the physical state are the most important parameters that characterize the morphology of the spectrum. This direct link can be numerically determined by radiative transfer models. Deducing the physical parameters from a spectrum cannot be solved analytically and requires the use of mathematical methods: optimization techniques, statistics, Bayesian approaches, variational methods or black-box models such as neural networks and support vector machines. From a general point of view, deducing some model parameters from the observed data is called an inverse problem. Inverse problems occur in many branches of sciences and mathematics such as geophysics, medical imaging, pedology, remote sensing, and astronomy... A synthesis on inverse problems can be found in [26], [27] and [19]. An interesting comparison between inversion methods can also be found in [14]. Very briefly, inverse problems can be described by the following equation:

$$x = G(y) \tag{1.2.1}$$

where y are the values of the model parameters, x the observed data and G is a linear or non linear operator, describing the explicit relationship between data and model parameters, and representing the physical system.
Deducing x from y knowing G is called a direct problem, whereas deducing y from x knowing G is called the inverse problem. In the case of hyperspectral images, x represents a spectrum and y the parameters in the radiative transfer model. Many approaches have been explored to solve such problems. In the domain of remote sensing, most of literature can be found in agronomy with problems such as the retrieval of canopy biophysical variables from reflectance, in oceanography for mapping bottom type, in pedology studying soil properties and in planetology for mapping physical parameters of planet surfaces. Coarsely summarizing the bibliography, methods can be classified in 3 categories:

- *Optimization algorithms.* These are the most traditional approaches. They require a physical model (here, radiative transfer model) able to simulate the spectra for different values of the parameters. They consist in minimizing over parameters a merit function expressing the similarity between an observed spectrum and a simulated spectrum. This function can be for example the mean square errors function where the errors are the differences between observed reflectances and simulated reflectances at each wavelength. These methods involve numerical optimization techniques (Powell's method, Simplex method, quasi-Newton method...) that start with an initial guess of the parameters and search for the optimum parameters thanks to an iterative process minimizing the merit function [14]. The problems of these approaches are that they are computationally heavy and time consuming because they simulate iteratively new spectra. In that way, they cannot be applied in case of the inversion of an image with numerous pixels. Moreover, they can sometimes fall into local minima and give consequently unstable solutions. Solution stability requires that the global minimizer does not change significantly for reasonable errors in the measured spectrum. In fact, instability is usually observed because inverse problems are often ill-posed (a small change in the data can lead to enormous differences in the estimations and solutions are not unique). A probabilistic approach (Monte Carlo approach) can then be used in inverse problems by introducing an a priori distribution on model parameters. In this approach, one considers that there is not only one solution of the optimization problem, but many solutions described by a probability density [19], [26] .

- *Look-up tables approach (LUT approach).* The idea of the look-up table approach is to replace an heavy runtime computation with a simpler look-up operation. A large database (look-up table) is generated by radiative transfer for many parameters values and stored in memory. Then, to reverse an hyperspectral image, the spectrum at each pixel is compared with the look-up table spectra in order to find the best match according to a merit function minimization. Parameters are then deduced from the look-up table best match spectrum. In comparison with traditional optimization methods, the speed gain is really significant, since retrieving a value from memory is often faster than undergoing an expensive computation. These methods are have been used in oceanography [20], [18] and in forestry studies [6]. The disadvantage of this approach is the instability of solutions due in particular to the non unicity of solutions. Many questions still remain on how to choose the merit function, how many spectra in the look up table have to be kept to estimate parameters, how to choose the look-up table, etc...

- *Training approach.* In the training approach, one makes the assumption that there exists a functional relationship f between spectra and parameters associating to each spectrum some parameters values (this relationship corresponds in fact to the inverse of the physical model G in the general inverse problem). The idea is to use the physical model to simulate spectra for a wide array of parameters values that constitute a learning database used to estimate the underlying mathematical relationship f. This relationship then allows to estimate the parameters of new spectra. The advantage of this approach is that once the relationship has been established, it can be used for very large sets and for all new images with the same physical model. Computation time can then be very competitive. Most of the time, neural networks are used to learn the underlying relationship between the set of input spectra and the set of output parameters [14], [6], [13], [29]. More recently, support vector machines (SVM) [4], [23], [25] propose a set of supervised learning methods used for regression. The basic idea is to map the data x into a high-dimensional feature space F via a nonlinear mapping, and to do linear regression in this space. These techniques have been used recently in remote sensing for solving the inversion problem of retrieving the leaf area index from imaging spectroradiometer [9].

In this report, we will discuss the LUT approach (also denoted by *k-nearest neighbors algorithm*) currently used by the Laboratoire de Planétologie de Grenoble to estimate physical properties y of Mars surface from spectral data x. The main purpose of this document is to propose a new approach based on a reduction dimension technique, the Sliced Inverse Regression, and on the functional estimation of f. Both methods require a learning database.

1.3 Data

1.3.1 Hyperspectral Images from Mars

In this report, four OMEGA hyperspectral images are analyzed. They have been acquired during orbits 30, 41, 61 and 103 that cover the high southern latitudes of Mars (see figure 1.3.2). The spatial resolution is about 2km per pixel and we considered 184 wavelengths in the range 0.95-4.15. For each image, a preprocessing aiming at correcting the atmospheric contribution in the spectra has been applied. For more details, see [8]. After treatment, these OMEGA observations revealed [2] that the south polar region of Mars mainly contains water ice, carbon dioxide ice and dust. A detailed qualitative mapping of H2O and CO2 ices during the local summer shows that the permanent south polar region is dominated by superficial CO2 on the bright cap except at its edges where water ice appears in extended areas. Examining the coexistence modes (geographical or granular) between H20, CO2 and dust that best explain the morphology of the spectra has then led to the implementation of a physical modeling of individual spectra with a surface reflectance model. This model allows the generation of synthetic spectra with the corresponding pairs of parameters that constitute a learning database [8]. In this report, we will not work on the whole images to reverse the model because of the diversity of physical models needed to simulate the whole image. We will focus on the terrain unit of strong concentration of CO2: the bright permanent south polar cap. This unit been determined by a classification method based on wavanglets developed at the Laboratoire de Planétologie de Grenoble [22]. For each image, the CO2 areas contain about 10000 to 20000 spectra.

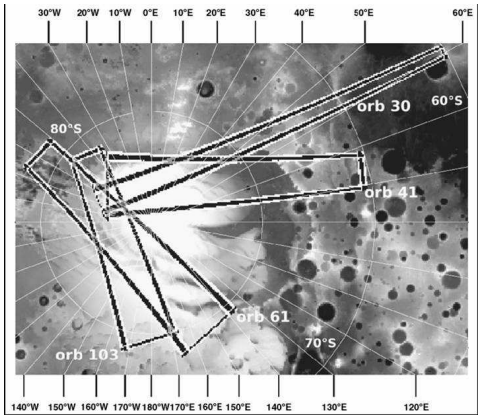


Figure 1.3.2: OMEGA hyperspectral images during orbits 30, 41, 61 and 103.

1.3.2 Learning databases

In order to estimate proportions and grain sizes of CO2, H2O and dust from OMEGA spectral images, two databases have been simulated for a range of parameters values judged representative of the south polar region. These two databases are derived from a physical model representing a granular mixture of three chemical elements: C02, H2O and dust. The spectrometer angle, the grain size of dust and some others parameters have been fixed to a constant value and five parameters are considered to vary spatially: the proportion of water, the proportion of CO2, the proportion of dust, the grain size of CO2 and the grain size of water. In fact, only four parameters are studied because the sum of the three proportions is equal to 1.

The first database (denoted Ldata 1) contains a smaller range of values than the second database (denoted Ldata 2). See table 1.1 for details. In fact, Ldata 2 has been simulated after Ldata 1 because Ldata 1 did not contain a sufficient range of values for most of the parameters. For example, using K-nn with Ldata 1:

- the grain size of CO2 is estimated to be the maximum value for 24% of the pixels,
- the grain size of water is estimated to be the maximum value for 90% of the pixels,
- the proportion of CO2 is estimated to be the maximum value for 30% of the pixels,
- etc...

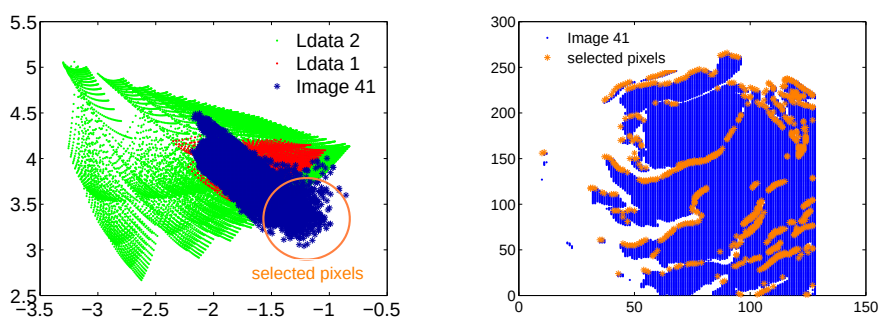
It seemed necessary to widen the learning database. Using K-nn with Ldata 2, then the number of reached maxima is reduced:

- the grain size of CO2 is estimated to be the maximum value for 4% of the pixels,
- the grain size of water is estimated to be the maximum value for 20% of the pixels,
- the proportion of CO2 is estimated to be the maximum value for 7% of the pixels,
- etc...

	Ldata 1 (*3584 spectra*)		Ldata 2 (*31500 spectra*)	
parameters	range	# distinct values	range	# distinct values
Prop. of water	[0.0006 0.002]	8	[0.0001 0.0029]	15
Prop. of CO2	[0.996 0.9988]	15	[0.9942 0.9998]	29
Prop. of dust	[0.0006 0.002]	8	[0.0001 0.0029]	15
Grain size water	[100 400]	4	[50 450]	5
Grain size CO2	[40000 105000]	14	[30000 165000]	28

Table 1.1: Sampling strategy for the learning databases

We kept both learning databases for the study because it could be helpful to understand how to choose a learning database and to better understand the methods. In order to visualize the adequacy of the two learning databases for the study of images from the south polar cap of Mars, we used a principal component analysis (see appendix B). We applied principal component analysis on the learning database Ldata 2. The projections of Ldata 1, Ldata 2 and of the real image from Mars during orbit 41 (denoted **image 41**) on the plane composed by the first and second PCA axis are presented figure 1.3.3 A. This figure shows that Ldata 1 projections do not cover entirely image 41 projections whereas Ldata 2 projections are spread out over a rather large area around image 41 projections. Moreover, for both databases, there exists an area for which image 41 projections are not covered. We denote this area in the graphics by the expression **selected pixels**. As we can see in figure 1.3.3 B, these pixels correspond to the boundary of the CO2 area. We can interpret these results as the fact that at the boundary of the CO2 bright area, the chosen physical model is not valid anymore. In the future work, estimations at the cap boundary should be interpreted with care or removed.



A B

Figure 1.3.3: A: Projections of the learning databases and the image 41 on the first PCA axes (of the spectra from image 41). Horizontally: First PCA factor. Vertically: Second PCA factor. B: Position of the **selected pixels** on the south polar cap map. Horizontally: x-coordinate. Vertically: y-coordinate.

In this report, we propose a general methodology to reverse hyperspectral images using a regularized version of Sliced Inverse Regression. This methodology is first validated on simulations and then applied on real datasets from Mars Express Mission.

In the first chapter, we present the current inversion used by the Laboratoire de planétologie de Grenoble based on a Look-up table simulation (LUT) and a K-nearest neighbors algorithm. We then propose a methodology based on the estimation, from a learning database, of a functional relationship between parameters and spectra. Because of the curse of dimensionality, a regularized Sliced Inverse Regression (GRSIR) is proposed to choose a lower dimensional regressor space sufficient enough to describe the relationship. Combined to a linear interpolation technique, this method allows to retrieve easily and quickly the parameters that generated some observed spectra.
In the second chapter, we validate the GRSIR methodology on simulations and compare it with the LUT approach. Some validation criteria are proposed and the choice of the different parameters introduced in the methodology are discussed. A methodology to choose the learning database is also developed.
Finally, in the third chapter we present the inversion of the four OMEGA images from the south polar car of Mars and discuss the realism of these results even if ideally some ground measurement are required.

Chapter 2

Methodology

In this chapter, we first present the LUT approach currently used by the Laboratoire de Planétologie de Grenoble and its limits (section 2.1). We then develop in section 2.2 a new methodology to reverse hyperspectral images, based on Sliced Inverse Regression. It consists in estimating the functional relationship between the spectra and the model parameters from a simulated learning database. Because of the curse of dimensionality, a Sliced Inverse Regression is considered in order to reduce the dimension of the data. A Regularized Sliced Inverse Regression (GRSIR) is finally proposed because of the ill-posedness of inverse problems. It aims at making the estimations more stable or smooth incorporating some prior information. Finally, a simple linear interpolation technique is used in section 2.2.3 to estimate the relationship between reduced spectra and parameters.

2.1 Current approach: Nearest neighbors algorithm

Let $x = (x^1, \ldots, x^d)$ be the $d = 184$ reflectances of one observed pixel on Mars.
Let $b^i = (b_i^1, \ldots, b_i^d), i \in \{1, \ldots, N\}$ be the d reflectances of the ith spectrum of the N simulated spectra of the learning database.
The nearest neighbor algorithm (also known as the K-nn algorithm) consists in searching the K nearest spectra of x in the learning database minimizing the mean square errors between the observed spectrum and the simulated spectra. The K nearest spectra of the learning database are selected sorting in ascending order the merit function $\Phi(i)$ in equation (2.1.1) with $i \in \{1, \ldots, N\}$:

$$\Phi(i) = \sum_{j=1}^{d} (x^j - b_i^j)^2. \tag{2.1.1}$$

If $K = 1$, the estimated parameters for the pixel are then the ones associated to the nearest spectrum selected in the learning database. If K is greater than 1, then one can choose to estimate parameters by the mode, or the average of the K best matches, but also to keep all the possible estimations. Most of the time, only the best match is retained.
In order to take into account the fact that some of the wavelengths can carry much more information about parameters, a weighted version of the nearest neighbor algorithm (denoted

WK-nn) has been proposed. It consists in assigning weights $p = p_1, \ldots, p_d$ to the most important wavelengths which transforms the merit function given in equation (2.1.1) into the merit function given in equation (2.1.2)

$$\Phi(i) = \sum_{j=1}^{d} p_j (x^j - b_i^j)^2. \tag{2.1.2}$$

The difficulty of WK-nn is to determine the weights to assign. An example is showed in figure 2.1.1. The same weights are assigned whatever the studied parameter is. In this case, we can observe that the nearest neighbor and the weighted neighbor are very close especially where weights are strong. Consequently, they lead to different estimations of the parameters (see table 2.1). One can see that for some parameters (proportion of CO2 and dust), WK-nn gives better estimations and for some others (grain size of CO2), K-nn gives better estimations.

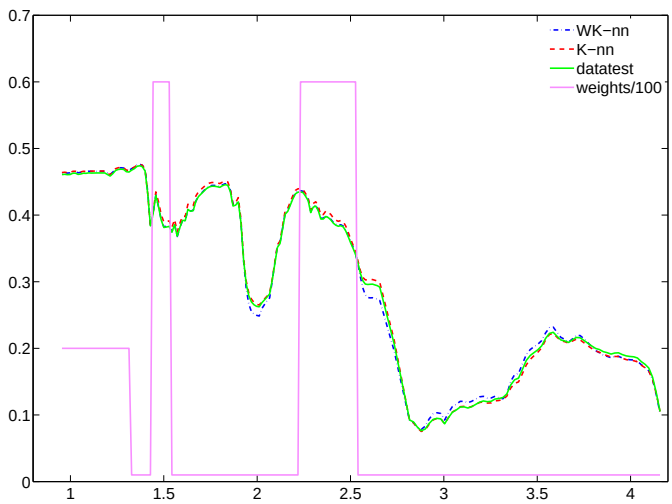


Figure 2.1.1: Example of comparison between the nearest neighbor algorithm and the weighted nearest algorithm. Learning database: Ldata 1. Spectrum 2888 from the test data Tdata (see chapter 3)

The problem of K-nn approach is that it generally leads to very unstable estimations and it is then difficult to choose a judicious K. Let us take the example of two spectra S1 and S2 for which parameters are known and let us add a reasonable noise to these spectra. Applying the K-nn methodology (K=1) to spectra with Ldata 2 leads to estimations given in table 2.2. We can notice here that estimation errors are relatively small for spectrum S1 whereas they are much greater for spectrum S2. One first thinks that a good idea to improve results would be to consider more than 1 neighbor. In fact, figure 2.1.2 shows how unstable the estimations are. It presents the relative estimation error as a function of K, where K is the Kth neighbor. First, we can see that two neighbors generally lead to very

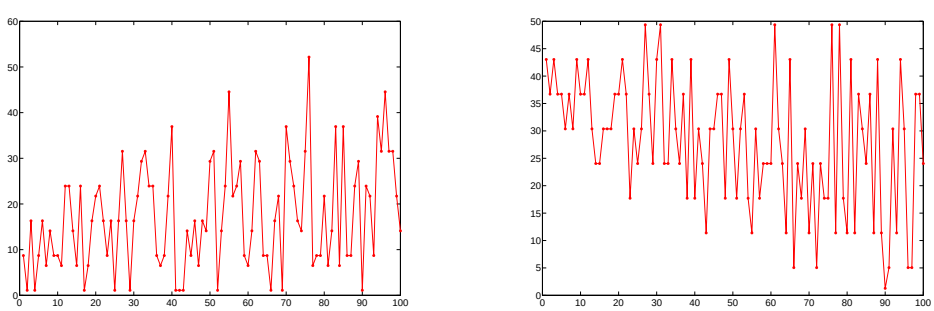
Parameters	True value	K-nn	WK-nn
Proportion of water	0.0011	0.0008	0.0014
Proportion of CO2	0.9971	0.9978	0.9966
Proportion of dust	0.0018	0.0014	0.0020
Grain size of water	157.5	100	200
Grain size of CO2	91151	105000	65000

Table 2.1: Example of comparison between the nearest neighbor algorithm and the weighted nearest algorithm. Learning database: Ldata 1. Spectrum 2888 from the test data Tdata (see chapter 3).

different estimations. Moreover, for spectrum S1, errors are increasing when K is increasing whereas for spectrum S2 errors are decreasing when K is increasing. That means that for spectrum S1, very few neighbors would be necessary to estimate the parameters properly whereas with spectrum S2, a lot of neighbors would be required. In fact for spectrum S2, more then one hundred neighbors are necessary to have at least a realistic estimation of the grain size of CO2. In these conditions, one can see how difficult it is to deduce parameters from spectra with K-nn methodology.

	Spectrum S1		Spectrum S2	
Parameters	Real values	K-nn	Real values	K-nn
Proportion of water	0.0019	0.0021	0.0013	0.0029
Proportion of CO2	0.9969	0.9966	0.9969	0.9942
Proportion of dust	0.0012	0.0013	0.0017	0.0029
Grain size of water	156	150	109	150
Grain size of CO2	65721	60000	78993	45000

Table 2.2: K-nn ($K = 1$) estimations for two test spectra (1600th and 2064th from datatest)



S1 S2

Figure 2.1.2: Relative errors on the grain size of CO2 for the 100 nearest neighbors in Ldata 2 of spectra S1 and S2. Horizontally:kth neighbor. Vertically: Relative errors on the grain size of CO2.

2.2 Proposed approach: Sliced Inverse Regression

In this report, we propose to establish a functional relationship between the spectra $X \in \mathbb{R}^d$ and different physical parameters $Y \in \mathbb{R}^p$. We want to estimate from a learning database, a function f, associating to each spectrum X some parameters values Y:

$$\begin{aligned} f : \mathbb{R}^d &\rightarrow \mathbb{R}^p \\ X &\mapsto Y = f(X). \end{aligned}$$

This relationship f corresponds in fact to the inverse of G in the general inverse problem (see equation (1.2.1)). However, estimating this relationship from the data is not realistic because of the curse of dimensionality. Indeed, it is impossible to accurately estimate the underlying regression function because the dimension of the regressor is too high and a huge amount of data would be required to fill the space densely. As a remedy, various dimension reduction techniques have been proposed to choose a lower dimensional regressor space that could be sufficient to describe the relationship of interest. The idea is then to find projection axes $\beta_1, \ldots, \beta_K$ with $K < d$ and $\beta_i \in \mathbb{R}^d$ for which there exists a function l from $\mathbb{R}^K$ to $\mathbb{R}^d$ such that:

$$Y = l(\beta_1 X, \ldots, \beta_K X) + \epsilon. \tag{2.2.3}$$

Principal component analysis is one of the most famous dimension reduction technique but applied to hyperspectral images, it does not seem sufficient enough to highlight the relationship between parameters and spectra. Results are presented in appendix B. Applying PCA to the learning database consists in applying PCA to the spectra of the learning database. Then one tries to show that a relationship exists between spectra projected on the first axes and parameters. In fact, in the dimension reduction step, only spectra are considered and parameters values are not taken into account. To take these values into account we propose to use Sliced Inverse Regression.

2.2.1 Sliced Inverse Regression

Sliced Inverse Regression (SIR) is a dimension reduction method developed by Li [15]. The idea of SIR is to consider that in a regression problem

$$Y = f(X) + \epsilon \tag{2.2.4}$$

one assumes it is efficient to consider a lower dimensional space $\beta = (\beta_1, ...\beta_K)$ such that there exists a function g from $\mathbb{R}^K$ to $\mathbb{R}^p$:

$$Y = g(\beta_1 X, \ldots, \beta_K X) + \epsilon \tag{2.2.5}$$

where:

- $Y = (Y_1, \ldots, Y_p)$ denotes the response variable,
- $X = (X_1, \ldots, X_d)$ is a d-dimensional random vector of explanatory variables with expectation μ and covariance matrix Σ,
- $\beta_1, \ldots, \beta_K$ are d-dimensional vectors with $K < d$,

- ϵ is a random error independent of X
- g is an unknown functional parameter.

In this report, we will consider that the response variable is defined in $\mathbb{R}$ ($p = 1$). Each parameter will be studied individually.
The aim of SIR is to estimate the subspace $E = Span(\beta_1, \ldots, \beta_K)$ of $\mathbb{R}^d$ called the effective dimension reduction space (the EDR-space) or equivalently the unknown β_i's called the effective dimension reduction directions (EDR-directions).
Let us denote by Z the standardized version of X. It has been shown by [15] that given the model (2.2.5), for any monotonic function T and under the condition that for all $b \in \mathbb{R}^p$, there exist $c_0, \ldots, c_K$ such that:

$$E[b^t X | \beta_1^t X, \ldots, \beta_K^t X] = c_0 + c_1 \beta_1^t X + \ldots + c_1 \beta_K^t X, \tag{2.2.6}$$

the covariance matrix $\Gamma_Z = Cov(E(Z|T(Y)))$ of the regression curve $Y \mapsto E(Z|T(Y))$ is degenerated in each direction orthogonal to all EDR directions. The consequence is that EDR directions can be deduced from the eigenvectors associated to the K higher eigenvalues of Γ_Z. As a consequence, they can also be deduced from the eigenvectors associated to the K higher eigenvalues of $\Sigma^{-1}\Gamma$ where $\Sigma = cov(X)$ is assumed to be regular and where $\Gamma = Cov(E(X|T(Y)))$.

The condition given by equation (2.2.6) can also be replaced by a stronger hypothesis that X has an elliptically symmetric distribution. Both conditions are difficult to verify in practice, but it has been shown that they could be be admitted when the dimension of X is high [12].

In applications, only a sample $(x_i, y_i), i = 1, \ldots, n$ is available, where n is the sample size, x_i is a d-dimensional data and y_i is a real number. In order to easily estimate Γ matrix, function T is chosen as the discretization of Y into slices. Thus, SIR computation requires the 4 following steps:

- Step 1: Sorting Y in increasing order and divide it into H slices S_h, $h = 1, \ldots, H$. SIR methodology generally requires the choice of the number of slices H. Here, this choice will be determined by the sampling strategy of the learning databases (see section 3.4).
- Step 2: Computing the slice means. For each slice S_h:

$$\hat{m}_h = \frac{1}{n\hat{p}_h} \sum_{i=1}^{n} (x_i - \bar{x}) \mathbb{I}_{y_i \in S_h} \text{ with } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.2.7}$$

  where

$$\hat{p}_h = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{y_i \in S_h} \tag{2.2.8}$$

  denotes the proportion of observations in S_h.

- Step 3: Computing the "between slices" means covariance matrix:

$$\hat{\Gamma} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h{}^t \tag{2.2.9}$$

- Step 4: Estimating the EDR $\beta_1, \ldots, \beta_K$ computing the eigenvectors of $\hat{\Sigma}^{-1}\hat{\Gamma}$ with

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t \tag{2.2.10}$$

An example of Sliced Inverse Regression is presented in figure 2.2.3 showing the projection of the data X on the first SIR axis as a function of the parameter Y. The interest of this curve is to visualize the fact that SIR method aims at maximizing the between slice variance of projections. It is easy to see that the relationship between the parameter and the projections on the EDR subspace is better when the between-slice variance is high. Equivalently, the relationship is excellent when the within-slice variance is close to zero.
Let us notice that the first direction β_1 can be viewed as a vector of weights and could be used for example in the weighted version of K-nn (see section 2.1).

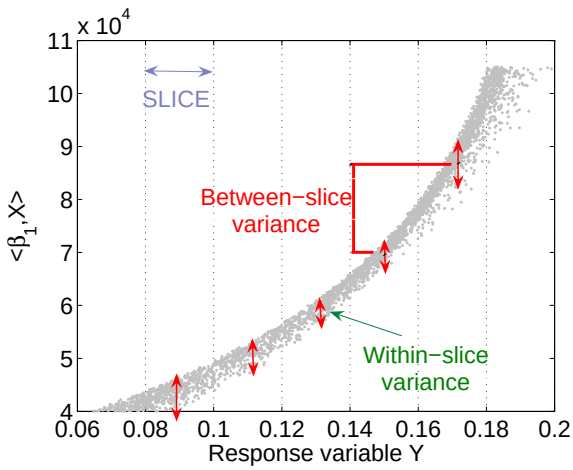
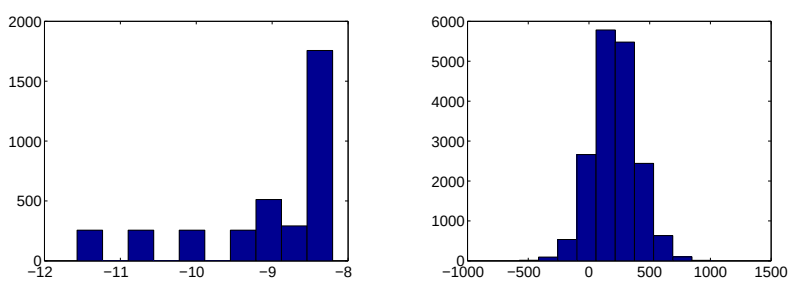


Figure 2.2.3: Projection on the first SIR axis of the Inverse Regression curve. Horizontally: Parameter y. Vertically: Projection of the data x on the first SIR axis.



Ldata 1 Image 41

Figure 2.2.4: Histograms of the projections of the spectra from Ldata 1 and the image from orbit 41 on the first SIR axis determined by a SIR analysis of Ldata 1.

When we applied SIR to hyperspectral image Ldata 1, obvious relationships appeared between parameters and projections of the data on the first SIR axis. These relationships are presented in figure 2.2.5. Results were very satisfying at first, but in fact, when we projected the hyperspectral image from orbit 41 on these axes, we could observe that the projections were varying in a completely different range of values (fig. 2.2.4). In fact a more thorough analysis showed that a small amount of noise in the data could lead to enormous differences in the histogram of the projections. This observation is very common in inverse problems because they are often ill-posed. Generally, ill-posed problems can be solved numerically introducing regularization techniques. The concept of these methods is to incorporate some prior information on the solution in order to damps the effect of the noise in the input data and to make the solution more regular or smooth. The ill-posed problem is then replaced by a slightly perturbed well-posed problem that depends on a parameter, called the regularization parameter, that ought to converge to zero in well-posed problems.

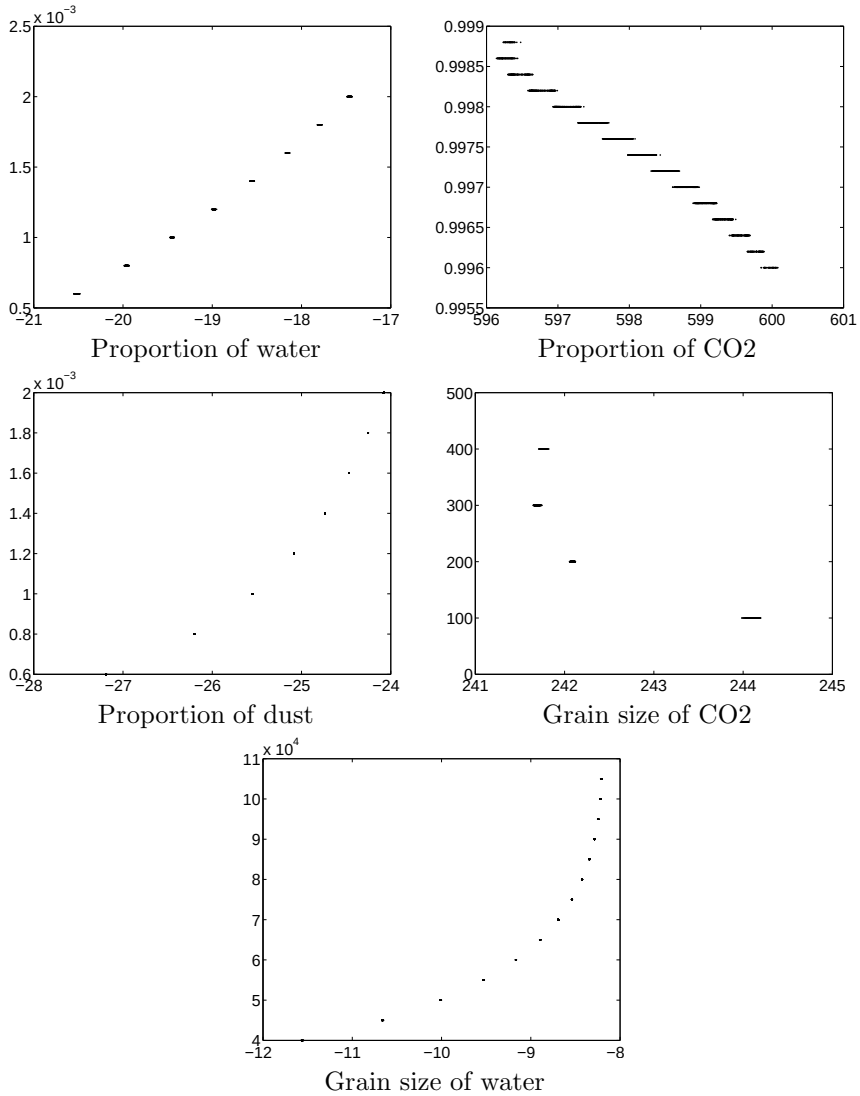


Figure 2.2.5: Different physical parameters values are presented as functions of the projections of spectra (from Ldata 1) on the first SIR axis

2.2.2 Regularized Sliced Inverse Regression

In the case of the application of SIR to planetary data, the necessity of regularization comes from the large condition number of the covariance matrix $\hat{\Sigma}$ of the spectra X. Indeed, in the SIR methodology, the central subspace is obtained by computing the largest eigenvalues of $\hat{\Sigma}^{-1}\hat{\Gamma}$ where the condition number of $\hat{\Sigma}$ is around 10^{14}! Inverting $\hat{\Sigma}$ in Sliced Inverse Regression generates then a strong instability in the estimation of the EDR space. From a general point of view, inverting a singular or ill-conditioned matrix A can be seen as solving in v an ill-conditioned linear system:

$$u = Av \tag{2.2.11}$$

where:

- u is a $m \times 1$ vector,
- v is a $n \times 1$ vector,
- A is a $m \times n$ matrix.

According to Hadamard [11], such a system is said "well posed" if:

- it has a solution,
- the solution is unique,
- the solution depends continuously on data and parameters (in some appropriate norms, small changes in data and parameters result in small changes in solutions).

Conversely, a system is "ill posed" if it is not "well posed". When a problem is ill-posed, the consequence is that a small amount of noise in the data can lead to large changes in the solution. In that case, one should use regularization techniques to compute more accurate solutions. Tikhonov regularization [28] is the most common method of regularization. It relies on the classical resolution of the linear system (2.2.11) by minimizing the functional:

$$\Phi(v) = \| Av - u \|_2^2 \tag{2.2.12}$$

that gives the well known estimator $\hat{v}$ of v

$$\hat{v} = (A^t A)^{-1} A^t u. \tag{2.2.13}$$

Since in ill-posed systems, such an estimator is very unstable, the proposed solution by Tikhonov is to add a term to $\Phi(v)$ penalizing the large components. It makes a compromise between minimizing the norm of the residuals $Av - u$ and minimizing the norm of the solution v. Thus instead of minimizing the functional given in equation (2.2.12), one minimizes:

$$\Phi(v) = \| Av - u \|_2^2 + \delta \| Mv \|_2^2 \tag{2.2.14}$$

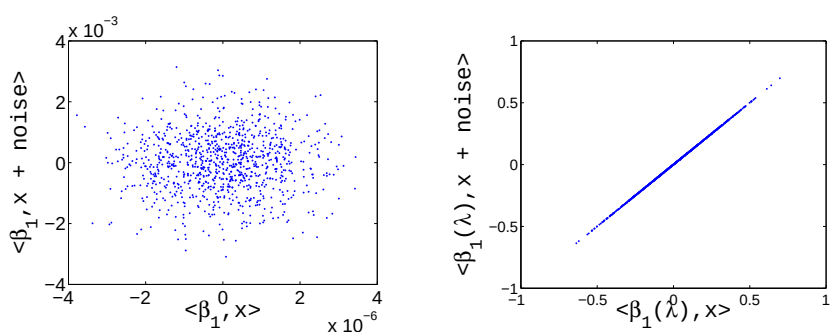
where M is a differential operator chosen according to the desired properties for the solution and δ is the regularization parameter. This way of proceeding improves the conditioning of the problem and enables a numerical solution. The explicit solution is then given by:

$$\hat{v} = (A^t A + \delta M^t M)^{-1} A^t u \tag{2.2.15}$$

When M is the identity operator, Tikhonov regularization is also well known as the ridge regression. For $\delta = 0$, Tikhonov regularization leads to the least squares solution (2.2.13).

Regularization techniques in linear or non linear regression problems are very famous and often used, but they only have been recently introduced in Sliced Inverse Regression. Different regularizations of the SIR method have been proposed. In [5] and [16], a principal component analysis is used as a preprocessing step in order to eliminate the directions in which the spectra are degenerated. Thus, for a properly chosen dimension of the projection subspace, the covariance matrix of the projected observations is regular. In [30], the sample estimate $\hat{\Sigma}$ is replaced by a perturbed version $\hat{\Sigma} + \delta I_d$ where I_d is the $d \times d$ identity matrix and δ a positive real number. Similarly, in [24], regularized discriminant analysis [10] is adapted to the SIR framework. More recently, it is proposed in [17] to interpret SIR as an optimization problem and to introduce L_1- and L_2- penalty terms in the optimized criterion. Our approach [1] is based on a Fisher Lecture given by R.D. Cook [7] where it is shown that the axes spanning the central subspace can be interpreted as the solution of an Inverse Regression problem. Details can be found in [1], but in this report, we will merely describe the idea of our methodology. In SIR, the main problem is to invert the covariance matrix $\hat{\Sigma}$ when computing the eigenvectors associated to the eigenvalues of $\hat{\Sigma}^{-1}\hat{\Gamma}$. Because $\hat{\Sigma}$ is most of the time singular or ill conditioned in the context of inverse problems, the first idea is then to replace $\hat{\Sigma}^{-1}$ by a more stable inverse $\widetilde{\Sigma}^{-1}$ such as pseudo inverse, or Tikhonov inverse. Then, Regularized Sliced Inverse Regression consists in computing the eigenvectors associated to the K higher eigenvalues of $\widetilde{\Sigma}^{-1}\hat{\Gamma}$. As in [1], we will refer to this regularized version of SIR by the Gaussian Regularized Sliced Inverse methodology (GRSIR). In this report, two regularizations will be tested:

- $\widetilde{\Sigma}^{-1} = (\Sigma + \delta I_d)^{-1}$ where δ is a positive real number called "regularization parameter". It corresponds to the regularized SIR developed by Zhong et al. [30]. We will refer to this methodology by "Zhong GRSIR".
- $\widetilde{\Sigma}^{-1} = (\Sigma^2 + \delta I_d)^{-1}\Sigma$. We will refer to this methodology by "Tikhonov GRSIR".



A: No regularization B: With Tikhonov regularization

Figure 2.2.6: Influence of the regularization on SIR first axis β_1. Comparisons of the projections of spectra from Ldata 1 on this axis β_1 with the projections of the same data with added noise. Horizontally: Projections of the spectra from Ldata 1 on the first SIR axis. Vertically: Projections of the disturbed spectra from Ldata 1 on the first SIR axis.

An example of Tikhonov GRSIR application is presented on figure 2.2.6. We considered for this application the learning database Ldata 1 and a noisy version of Ldata 1 called N-Ldata 1 where we added a Gaussian noise to all spectra. We applied SIR (2.2.6 A) and Tikhonov GRSIR (2.2.6 B) to Ldata 1 and we compared projections of the spectra from Ldata 1 and N-Ldata 1 on the first SIR axis for both methods. We can see that without any regularizations, projections on the first SIR axis (and consequently estimations) are very sensitive to the noise whereas with Tikhonov regularization they are not.

2.2.3 Estimation of the functional relationship

Once the relationship between parameters and projected spectra has been revealed, then the question is to estimate such a relation. Let us call $\beta_1(\delta)$ the first GRSIR axis with a regularization parameter δ. From a learning database, the question is now to estimate a function f from $\mathbb{R}$ to $\mathbb{R}$ associating to each new projected spectrum $< \beta_1(\delta), X >$ a parameter value Y. Many methods can be used to solve such a problem. We could use again the K nearest algorithm. Given a new spectrum X and its projection $< \beta_1(\delta), X >$ on the first SIR axis $\beta_1(\delta)$ determined with a learning database "Ldata", we look at the K nearest projections in $< \beta_1(\delta), Ldata >$ and their associated parameters. The estimated parameter value is then the average of these K parameter values. The problem of this method is that it is computationally heavy and time consuming. Spline interpolations can also be computed easily but they require the use of new parameters such as the number of nodes, the degree of the polynomials and moreover, they lead to very unstable estimations at the boundaries. That is why we propose to use a simple linear interpolation on the set of data points $(m_h^{proj}, m_h^{param}), h = 1, \ldots, H$ where H denotes the number of slices S_h, m_h^{proj} denotes the average projection of spectra for slice h and m_h^{param} denotes the average parameter value for slice h.

$$m_h^{proj} =< \hat{m}_h + \bar{x}, \beta_1(\delta) >, \tag{2.2.16}$$

and

$$m_h^{param} = \frac{1}{n\hat{p}_h} \sum_{i=1}^{n} y_i \mathbb{I}_{y_i \in S_h} \tag{2.2.17}$$

For each new spectrum X with a projection $< \beta_1, X >$, the estimated parameter value $\hat{Y}$ is then given by:

$$\begin{aligned}
\hat{Y} &= \hat{f}(< \beta_1(\delta), X >) \\
&= m_1^{param} \mathbb{I}_{<\beta_1(\delta), X> \in]-\infty, m_1^{proj}]} + m_N^{param} \mathbb{I}_{<\beta_1(\delta), X> \in]m_N^{proj}, \infty[} \\
&+ \sum_{h=1}^{N-1} \left[m_h^{param} + (< \beta_1(\delta), X > - m_h^{proj}) \left(\frac{m_{h+1}^{param} - m_h^{param}}{m_{h+1}^{proj} - m_h^{proj}} \right) \right] \mathbb{I}_{<\beta_1(\delta), X> \in]m_h^{proj}, m_{h+1}^{proj}]}
\end{aligned}$$

An example of linear interpolation is given on figure 2.2.7. At the boundaries, estimators have been fixed to the minimum and maximum average values observed on the learning data base.

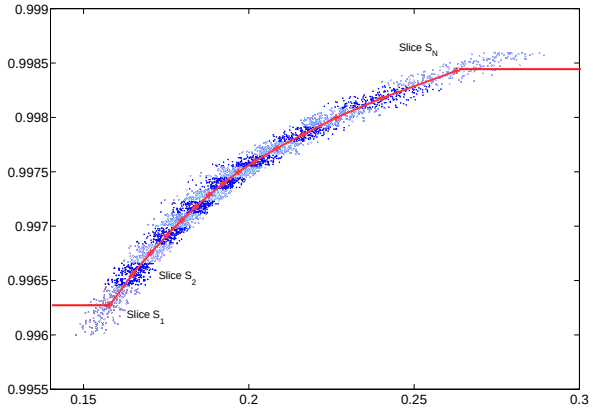


Figure 2.2.7: Estimation of the relationship between projections of spectra on the first SIR axis and one parameter by linear interpolation. Horizontally: projections of spectra on the first SIR axis. Vertically: proportion of CO2. Learning database: Tdata (see chapter 3) .Used methodology: Tikhonov GRSIR

Chapter 3

Validation on simulations

In this chapter, we compare and validate the different proposed methods mentioned previously: K-Nearest Neighbors (K-nn), Weighted K-Nearest Neighbors (WK-nn), Zhong and Tikhonov Gaussian Regularized Sliced Inverse Regression (Zhong and Tikhonov GRSIR).
In section 3.1, a test data is simulated by radiative transfer models for validation.
Then, two validation criteria are proposed in section 3.2: the Normalized Root Mean Square Errors criterion (NRMSE) quantifying the importance of errors and the Sliced Inverse Regression criterion (SIRC) quantifying the quality of the relationship between reduced spectra and parameters. The minimization of the NRMSE criterion is used in section 3.3 to choose the regularization parameter.
Then different steps of the GRSIR methodology are discussed:

- in section 3.4, we show that GRSIR is not strongly sensitive to the choice of the slices,
- in section 3.5, validation criteria are compared for the different proposed methods and results show that Tikhonov GRSIR seems to be the best one,
- in section 3.6, we propose a methodology based on Principal component analysis and Gaussian mixture models to select an appropriate learning database for the inversion,
- and because the sum of the proportions has to be one, we propose in section 3.7 to estimate only two of the proportions and to deduce the last one by subtracting the two estimated proportions from one.

Finally, we conclude this chapter with the results obtained by the GRSIR methodology taking into account all these previous steps.

3.1 Simulation of a test data

For the validation sake, the use of a test dataset is required. The one we chose has been simulated by radiative transfer modeling for random values of the following parameters: the proportion of water (varying from 0.0006 to 0.002), the proportion of CO2 (varying from 0.996 to 0.9988), the grain size of CO2 (varying from 40000 to 105000) and the grain size of

water (varying from 100 to 400). In fact these parameters are exactly varying in the same ranges of values as the learning database Ldata 1 but they have been chosen randomly. In order to work in realistic conditions, a multiGaussian noise of dimension 184 has been added to all the spectra of the test dataset. We will denote this new data by "Tdata". The noise has been simulated with a mean fixed to zero for all wavelengths and with a covariance matrix determined experimentally from the OMEGA image acquired during orbit 41. A small portion of the image, very homogeneous in terms of composition and physical properties, is chosen so that we can assume that much of the variability comes from the noise. The latter is then evaluated by a statistics after applying a shift difference on the selected portion.

Let us consider:

- n_T the number of spectra from Tdata,
- $x_i^T \in \mathbb{R}^{184}, i \in 1, \ldots, n_T$ the spectra from Tdata,
- $y_i^T \in \mathbb{R}, i \in 1, \ldots, n_T$ the associated values for one parameter y of Tdata,
- n_L the number of spectra from the learning database, the learning database being for example Ldata 1 or Ldata 2,
- $x_i^L \in \mathbb{R}^{184}, i \in 1, \ldots, n_L$ the spectra from the learning database,
- $y_i^L \in \mathbb{R}, i \in 1, \ldots, n_L$ the associated values for one parameter y of the learning database.

The main idea of validation is to estimate the parameters values $\hat{y}_i^T$ of the test data from a learning database using the different methods mentioned above and to compare these estimations $\hat{y}_i^T$ to the real values y_i^T.

3.2 Validation criteria

To quantify the advantages of one method compared to another, we had a look at two aspects of the validation: the quality of the estimations but also the quality of the relationship between spectra and parameters in the case of Sliced Inverse Regression. To this end, we developed two validation criteria. The first one, denoted SIR Criterion (SIRC) is the ratio between the "between-slices" variance of SIR projections and the total variance. It quantifies the quality of the relationship between projected spectra and parameters. The second one, denoted Normalized Root Mean Square Errors (NRMSE), quantifies the importance of estimation errors, that are the differences between the estimations and the real values. A third validation criterion that could also be used is the RMSE between the test spectrum x_i^T and the spectrum that can be reconstructed by running the radiative transfer model with the estimated parameters $\hat{y}_i^T$. Generally, this criterion will indicate that K-nn gives the best results which is logical because it aims at minimizing the RMSE. Moreover, this criterion faces the problem that two spectra can be very close in terms of RMSE even with very different parameters values. The consequences are that a good reconstruction does not always mean that parameters are well estimated. In the case of Zhong GRSIR and Tikhonov GRSIR, Gaussian Regularized Sliced Inverse Regression is applied as a first step to the learning database. Only the first SIR axis $\beta_1(\delta)$, depending on the regularization

parameter δ is considered because we will see later it is sufficient to explain the relationship between spectra and parameters in most of the cases.

3.2.1 Sliced Inverse Regression criterion

Applied to the learning database, the SIR criterion is defined as the ratio between the "between-slices" variance $\beta_1^t(\delta)\hat{\Gamma}\beta_1(\delta)$ of the projections of X_i^L on $\beta_1(\delta)$ and the total variance $\beta_1^t(\delta)\hat{\Sigma}\beta_1(\delta)$ of these same projections:

$$SIRC = \frac{\beta_1^t(\delta)\hat{\Gamma}\beta_1(\delta)}{\beta_1^t(\delta)\hat{\Sigma}\beta_1(\delta)} \tag{3.2.1}$$

This criterion indicates the quality of the functional relationship between projections of the spectra on the first SIR axis and parameters. As it was shown in figure 2.2.3, one can see that the relationship is more obvious and can be fitted more easily when the between slice variance is high or equivalently when the within slice variance is small. In the SIR criterion, the total variance can be interpreted as the sum of the between slices variance and the within slice variance. In this case, the quality of the relationship is perfect when the within variance is close to zero, or equivalently when SIRC is close to 1. Finally, the closer SIRC is from 1, the better the relationship is.

In GRSIR, the SIR criterion depends on the regularization parameter δ. Figure 3.2.1 shows that SIRC is strongly decreasing when the regularization parameter is increasing. Indeed, introducing regularization deteriorates the relationship between projected spectra and parameters as it is shown in figure 3.2.2. However, when no regularization is introduced, in the presence of noise, estimation errors are enormous. In fact, a compromise has to be made between deteriorating the relationship and improving estimations.

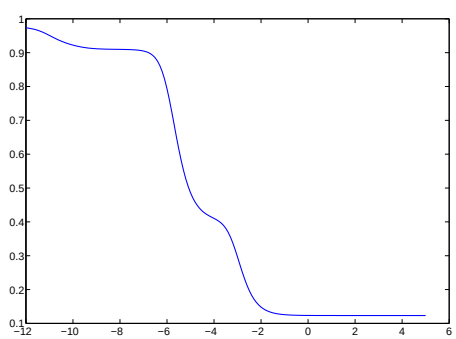
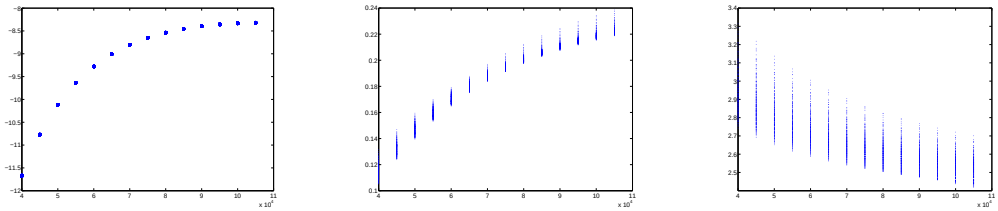


Figure 3.2.1: SIRC as a function of the regularization parameter. Horizontally: Regularization parameter. Vertically: SIRC. Learning database: Ldata 1. Studied parameter: grain size of CO2

3.2.2 Normalized RMSE criterion

In order to compare estimations to real parameters values, we decided to introduce the classical Root Mean Square Errors criterion that consists in calculating the root of mean



$\delta = 0$ $\delta = 10^{-10}$ $\delta = 10^{-4}$

$SIRC = 0.99$ $SIRC = 0.92$ $SIRC = 0.41$

Figure 3.2.2: Functional relationship between projections on the first SIR axis and the parameter "Grain size of CO2" for different regularization parameters. Horizontally: Grain size of CO2. Vertically: Projection of the spectra on the first SIR axis estimated with regularization parameters equal to 0, 10^{-10} and 10^{-4}. Learning database: Ldata 1.

square errors:

$$RMSE = \sqrt{\frac{1}{n_T}\sum_{i=1}^{n_T}(\hat{y_i}^T - y_i^T)^2}. \tag{3.2.2}$$

However, the RMSE criterion did not seem the most appropriate criterion to analyze results. Because all parameters are varying in very different ranges of values, it would not have been possible to compare the RMSE's between different parameters and to deduce if one parameter is generally better estimated than an other. That is why we introduced a normalized version of RMSE dividing it by the variance.

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n_T}(\hat{y_i}^T - y_i^T)^2}{\sum_{i=1}^{n_T}(y_i^T - \overline{y}^T)^2}}, \tag{3.2.3}$$

with

$$\overline{y}^T = \frac{1}{n_T}\sum_{i=1}^{n_T} y_i^T. \tag{3.2.4}$$

Obviously, estimations are better when the "normalized" RMSE, denoted "NRMSE", is close to zero.

3.3 Choice of the regularization parameter

In the case of Regularized Sliced Inverse Regression, estimations $\hat{y_i}^T$ and consequently NRMSE criterion as well as the SIRC score depend on a regularization parameter. As we saw before, when the regularization parameter increases the functional relationship between projected spectra and parameters is getting worse and consequently estimation errors

are increasing. From another point of view, if the regularization parameter is too small, then estimation errors are huge because we are dealing with an ill-posed problem. So there is a compromise to reach between deteriorating the functional relationship and improving estimations by regularizing. In fact, this compromise lies in the choice of the regularization parameter. We propose here to choose this regularization parameter, for each parameter individually, minimizing the NRMSE criterion calculated between the parameter values from the learning database $Y_1^L, \ldots, Y_{n_L}^L$ and their estimations $\hat{Y}_1^L, \ldots, \hat{Y}_{n_L}^L$. The latter are obtained by applying the estimated relationship $\hat{f}$ to the spectra from the learning database that has been spoiled by a multiGaussian noise. The relationship $\hat{f}$ has been calculated from the learning database using Regularized Sliced Inverse Regression. A synthesis about the methodology is given in figure 4.4.4. Here, the choice of the regularization is made applying the NRMSE criterion to the noisy learning database and not to the test data that has only be introduced for validation but not for the inversion of a real image. It is important here to insist on the fact that if there is no noise in the data, or in others words, if the observed data exactly corresponds to spectra that could be simulated by radiative transfer model, then no regularization is required and minimizing the NRMSE criterion for Regularized Sliced Inverse Regression would yield a value δ close to zero. In fact, the necessity of regularization comes from the fact that observed data always contain some noise and that in the case of ill posed problems, such a noise leads to enormous errors in estimations. The graphics presented in figure 3.3.3 show the evolution of the NRMSE criterion as a function of the regularization parameter for both Zhong and Tikhonov's GRSIR. We can see that considering one parameter, minima are approximately equal for both methods, whereas they are reached for a different range of values depending on the regularization method and the parameter. With Tikhonov regularization, it seems that the minimum is always reached on a larger range of values than with Zhong regularization. This point is interesting when applying GRSIR to real images, because in this case, the noise can be slightly different from the one introduced in the learning database and consequently, the chosen parameter should be slightly shifted. It is then better that a small shift of the regularization parameter does not lead to a big change in estimations. We insist here on the fact that the choice of the regularization is crucial for the inversion of the dataset because estimations can strongly vary when changing the regularization parameter value (see for example the inversions of image 41 for different parameters values in appendix E). This choice depends strongly on the noise. Figure 3.3.4 shows how the choice of the regularization parameter changes according to the noise:

- the stronger the noise is, the greater the NRMSE are,
- when the noise is increasing, the chosen regularization parameter is also increasing.

One can easily see that making a mistake on the estimation of the noise leads to a mistake in the estimation of the regularization parameter and consequently can lead to much more uncertain estimations.
In the next chapters and sections, results will always be presented for the optimum regularization parameter. The functional relationships between projected spectra and parameters are presented in appendix C and SIR weights are presented in appendix D.

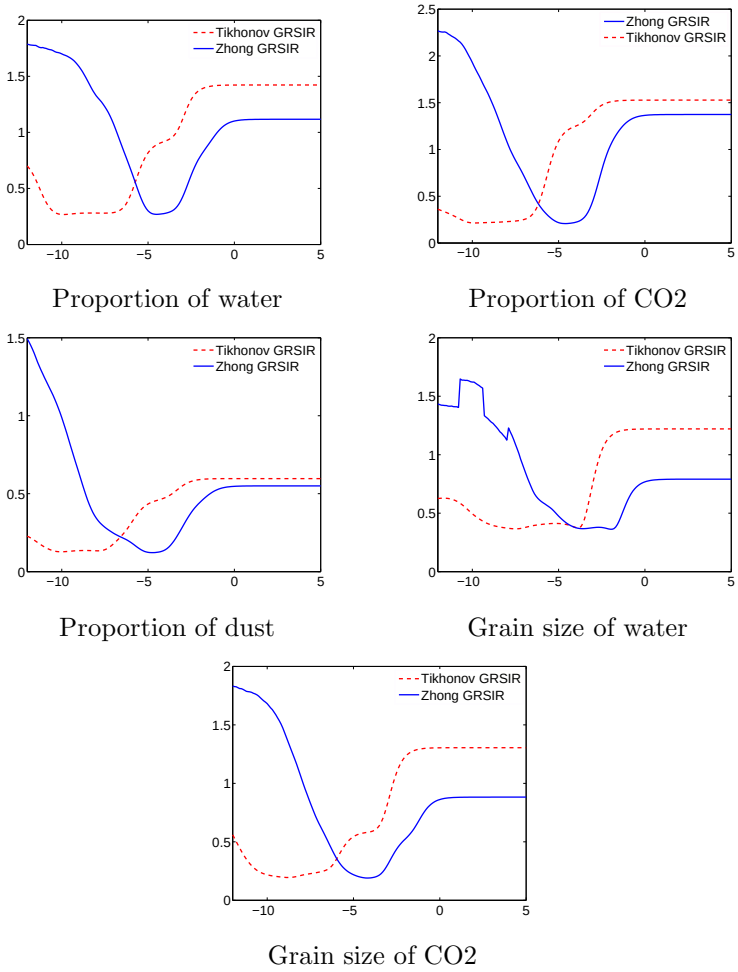


Figure 3.3.3: Evolution of the NRMSE criterion as a function of the regularization parameter for Tikhonov and Zhong's Gaussian Regularized Sliced Inverse Regression. Horizontally: Logarithm (to the base 10) of the regularization parameter. Vertically: NRMSE for Tikhonov and Zhong GRSIR. Learning database: Ldata 1

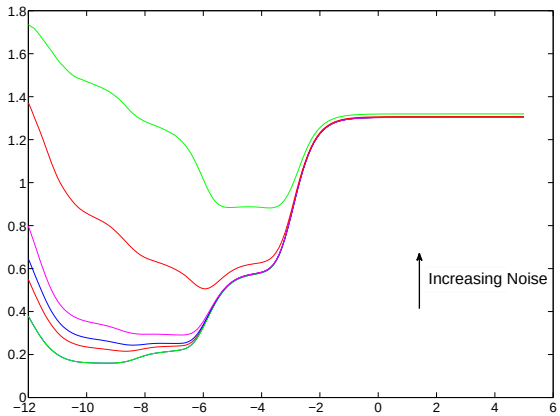


Figure 3.3.4: Influence of the noise in the choice of the regularization parameter. Horizontally: Regularization parameter. Vertically: NRMSE. A multiGaussian noise has been introduced in the data with a diagonal covariance C proportional to the identity matrix I_d: $C = \nu I_d$. The experience has been repeated 6 times with increasing values of ν. Learning database: Ldata 1. Studied parameter: grain size of CO2.

3.4 Influence of the slices

Gaussian regularized Sliced Inverse Regression depends on the choice of the slices. Generally, slices are chosen such that each slice contains the same number of observations but in our study, learning databases have been simulated for a fixed number of distinct parameters values and it could be more judicious to choose slices such that one slice corresponds to all spectra simulated for one fixed parameter value. For example, if spectra have been simulated for some grain sizes of CO2 equal to 40000, 50000 and 60000, the number of slices will be 3 and the first slice will contain all spectra simulated for a grain size equal to 40000, the second slice will contain all spectra simulated for a grain size equal to 50000 and so on... In that case, slices will not necessarily have the same size but they will be established according to the sampling strategy. In this report, this way of proceeding will be denoted by "sample slicing". In table 3.1, we analyze the influence of slicing on the SIRC and NRMSE when applying Tikhonov-GRSIR. Four slicing strategies are compared with Ldata 1: in the first column, slices are determined according to "sample slicing" and in other columns, data is divided into $nslices = 4, 20, 40$ slices of the same size. Results show that in most of the cases, SIRC and NRMSE are better for the "sample slicing" that we will use in all the next sessions. However, estimations and functional relationships are not strongly deteriorated using the others slicing strategies. In fact SIRC and NRMSE do not seem strongly sensitive to slicing.

Parameters	sample slicing		$nslices = 4$		$nslices = 20$		$nslices = 40$	
	NRMSE	SIRC	NRMSE	SIRC	NRMSE	SIRC	NRMSE	SIRC
Prop. of water	0.29	**0.92**	**0.28**	0.88	0.30	0.91	0.30	**0.92**
Prop. of CO2	**0.22**	**0.99**	0.28	0.82	0.25	0.97	0.26	0.98
Prop. of dust	**0.13**	**0.99**	0.16	0.93	0.19	**0.99**	0.18	**0.99**
Grain size water	**0.37**	**0.92**	**0.37**	**0.92**	0.39	**0.92**	0.43	0.91
Grain size CO2	**0.19**	0.98	0.21	0.90	0.20	0.98	0.20	**0.99**

Table 3.1: Influence of slices on NRMSE and SIRC. Used methodology: Tikhonov GRSIR. Learning database: Ldata 1.

3.5 Comparisons between methods

We present in this section the retrieval of all studied parameters for the four mentioned methods: K-nn, WK-nn, Tikhonov GRSIR and Zhong GRSIR, each one being evaluated by the validation criteria. The results presented in table 3.2 have been established with the learning database Ldata 1 and the ones presented in table 3.3 have been established with the learning database Ldata 2.
From these results, we can deduce that in most of the case, Tikhonov and Zhong GRSIR give similar validation criteria. In fact estimations are really close for both methods. The SIRC criterion is quite good for each parameter and for both methods but is deteriorated working with Ldata 2. The minimum NRMSE is reached for the proportion of dust and the maximum NRMSE is reached for the grain size of water (that is logical since the grain size of water has been sampled only for 4/5 distinct values in the learning databases). K-nn and WK-nn give worse estimations than the GRSIR methodology in any case. We

can also notice that WK-nn, introduced to improve estimations, deteriorates estimations for the grain size of water and the grain size of CO2. The comparison of the validation criteria established with Ldata 1 and Ldata 2 let appear that estimations are worse for all methods with the big database Ldata 2. This point will be discussed in section 3.6. In Figure 3.5.5, one can visualize estimations as a function of the real values for Ldata 1. We will not show these graphics with Ldata 2 because they are visually the same except that the variance of estimations is more important. Estimations of the proportion of dust can be surprising at first because they gather into some strata whereas they should be spread out randomly. In fact the proportion of dust, contrary to the other parameters, cannot be simulated completely at random since it is dependent on the proportion of CO2 and water (Proportion of dust = 1- proportion of water - proportion of CO2).
Because we saw in section 3.3 that Tikhonov GRSIR is more appropriate than Zhong GRSIR for the choice of the regularization parameter, we will only present results given by Tikhonov GRSIR in the next sections. For the sake of simplicity, we will denote the Tikhonov GRSIR methodology by GRSIR only.

Parameters	Tikhonov		Zhong		K-nn	WK-nn
	NRMSE	SIRC	NRMSE	SIRC	NRMSE	NRMSE
Proportion of water	**0.29**	**0.92**	**0.29**	**0.92**	0.50	0.38
Proportion of CO2	0.22	**0.99**	**0.2**1	0.98	0.54	0.46
Proportion of dust	0.13	**0.99**	**0.12**	**0.99**	0.34	0.35
Grain size of water	**0.37**	**0.92**	0.38	0.87	0.39	0.45
Grain size of CO2	0.19	**0.98**	**0.18**	**0.98**	0.35	0.46

Table 3.2: Estimation of the test data parameters by Ldata 1

Parameters	Tikhonov		Zhong		K-nn	WK-nn
	NRMSE	SIRC	NRMSE	SIRC	NRMSE	NRMSE
Proportion of water	**0.63**	**0.88**	**0.63**	0.87	0.86	0.60
Proportion of CO2	0.40	0.97	**0.38**	**0.98**	0.88	0.68
Proportion of dust	0.31	0.99	**0.28**	**0.99**	0.44	0.41
Grain size of water	**0.41**	0.80	0.47	**0.81**	0.43	0.48
Grain size of CO2	0.27	**0.93**	**0.26**	**0.93**	0.53	0.67

Table 3.3: Estimation of the test data parameters by Ldata 2

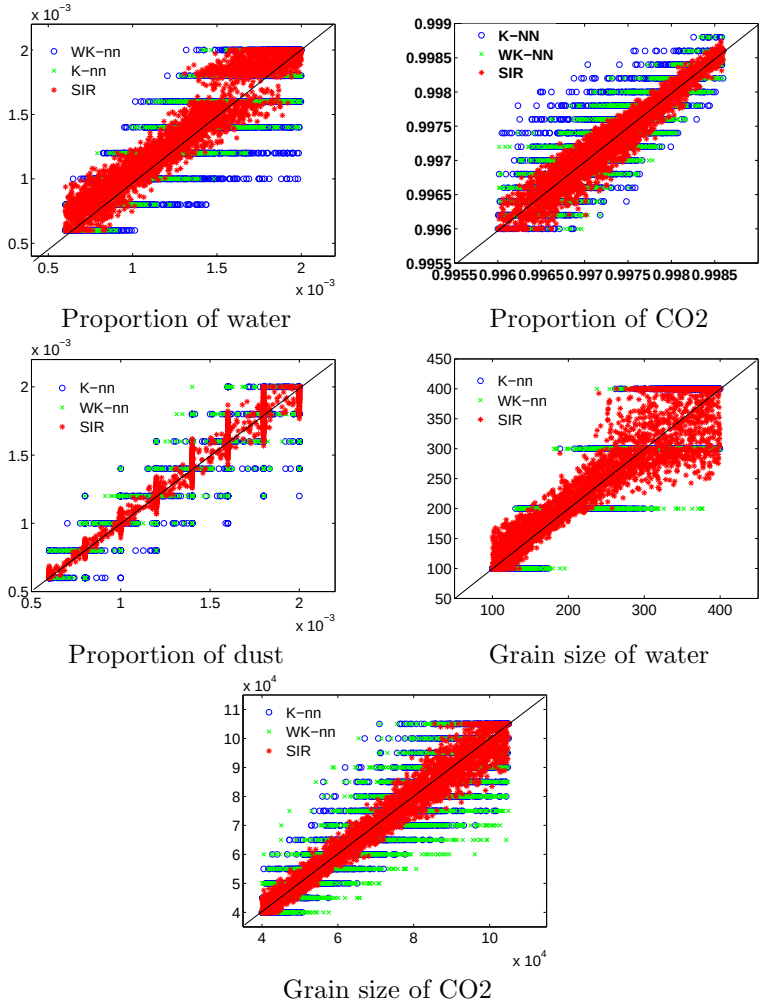
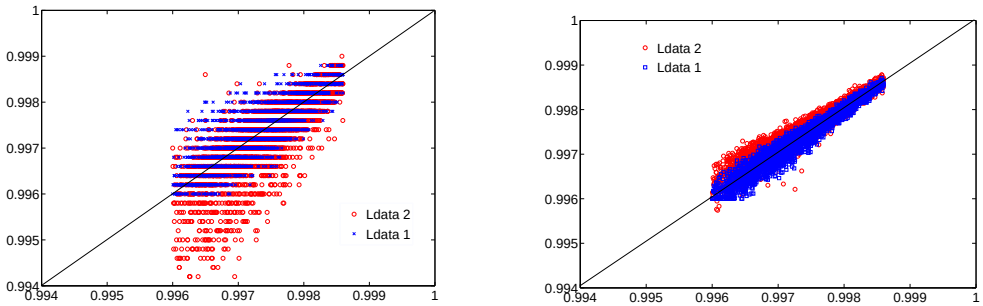


Figure 3.5.5: Scatter plot of the parameters values versus estimated values. Horizontally: Parameters values. Vertically: Estimated parameters values by GRSIR applied on Ldata 1

3.6 Choice of the learning database

The previous results given in table 3.2 and table 3.3 show that estimations of Tdata are deteriorated when using the learning database Ldata 2. This result is at first surprising because Ldata 2 has been simulated for much more distinct parameters values than Ldata 1. In fact Ldata 2 is approximately containing all the spectra from Ldata 1 so one could be surprised that adding some information would deteriorate estimations. Figure 3.6.6 A shows an example of the K-nn estimations of the proportion of CO2 with Ldata 1 and Ldata 2 as a function of the real parameters values from Tdata. We can observe that some of the proportions become underestimated with Ldata 2.
We found an explanation to these results. Let us consider for example the 2064th spectrum of the Tdata. The proportions of water, CO2 and dust for this spectrum are respectively 0.00134, 0.99695 and 0.0017 and the grain size of CO2 water and CO2 are 109.89 and 78993. Its nearest neighbor in Ldata 2 in terms of mean square errors is the 3811th spectrum with the following proportions of water, CO2 and dust: 0.0029, 0.9942, 0.0029 and grain sizes of water and CO2 equal to 150 and 45000. We can notice that this nearest spectrum has very different parameters values. In fact if we consider the 11619th spectrum of Ldata 2 with proportions and grain sizes much more closer to the ones associated to the 2064th spectrum (see figure 3.6.7 for values), we can see that this spectrum is more distant in terms of mean square errors (0.0266) than the 3811th spectrum (0.0130). This observation confirms that the inverse problem is ill-posed: on the one hand, we have seen that small noise in the data can lead to errors in estimations, and on the other hand, the solution of the inverse problem does not seem to be unique. These two conclusions are typical in ill-posed problems. The measurements are too limited to constrain a unique set of parameters in the framework of a given physical model for the surface. The inversion problem is then non uniquely determined, a situation that is widespread in remote sensing. If we had an a priori about the range of variations of each parameter, we could reduce the database and consequently reduce the number of possible solutions. But the problem is that we do not have any a priori information about the solutions and a judicious choice for the learning database is then difficult in case of K-nn methodology.
In figure 3.6.6 B, we can see that using Ldata 2 also deteriorates SIR estimations. But in Sliced Inverse Regression, the consequences of using Ldata 2 are different: proportions seem to be a little bit overestimated on the contrary to K-nn methodology. In fact, a more fastidious study shows that these overestimation are due to spectra distant from the test data whereas for K-nn, underestimations are on the contrary due to the addition of spectra close to the test data but with very different parameters values (an example can be found in figure 3.6.7).



A B

Figure 3.6.6: Estimations as a function of the true value. Comparisons between estimations realized with Ldata 1 and Ldata 2. A: KNN. B: GRSIR. Horizontally: True value . Vertically: estimated value. Studied parameter: proportion of CO2.

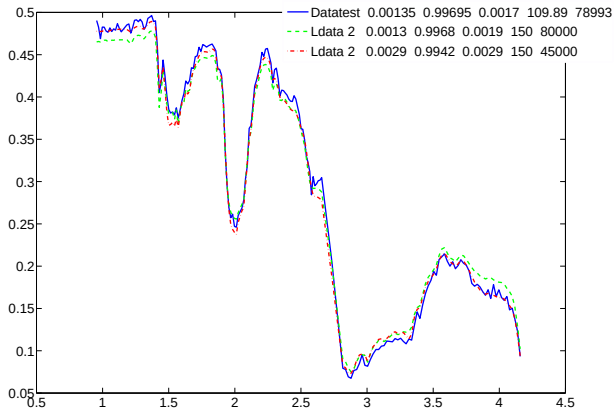
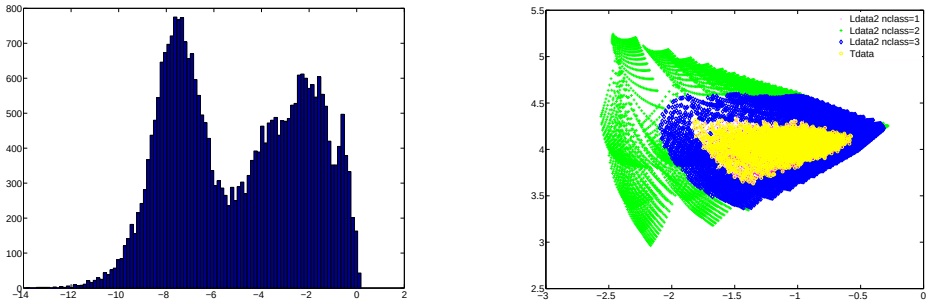


Figure 3.6.7: Analyzing the K-nn results for one spectrum of Tdata. Ldata 2. Horizontally: Wavelengths, Vertically: Reflectance

Let us consider the dataset Ldata 2. We know that in Ldata 2, many spectra are not necessary to estimate properly the parameters from Tdata. For example, if we remove from Ldata 2 all the spectra with parameters values that are not included in the range of values of Tdata's parameters, we obtain a new learning database denoted *Ldata 2 selection*, very similar to Ldata 1 that lead to better estimations than Ldata 2 for both K-nn and SIR (see table 3.4).
In practical cases, we do not have any a priori about the range of variations of the parameters. The best approach in that case is then to take a huge learning database and to reduce it after the first estimations by SIR or K-nn. But we have seen that when the database is too important, then estimations are strongly deteriorated. In the case of K-nn, these deteriorations are due to the similarity of some spectra with very different parameters. In the case of SIR, these deteriorations are mainly due to the spectra that are on the opposite different from the one observed. In we consider the projections (see figure 3.6.8 B) of Ldata 2 and Tdata on the first PCA axes deduced from the application of PCA to Ldata 2, we can see that a lot of spectra are not necessary to estimate Tdata's parameters. We propose in this document a methodology to select the most appropriate spectra for GRSIR. The idea is to retain the spectra from Ldata 2 whose projections on PCA axes are close to projections of the spectra from Tdata. It amounts to calculating, in the plane spanned by PCA axes, the closure of Tdata in Ldata 2. In order to calculate this closure, we calculate the distance of each spectrum from Ldata 2 with its nearest neighbor in Tdata. The histogram of these distances (see figure 3.6.8 A) allows to distinguish a mixture of 3 Gaussian densities. The application of EM algorithm allows the estimation of the parameters of this Gaussian mixture model (GMM) and the Maximum a posteriori classification (see [3] for details) divides spectra in three classes shown in figure 3.6.8 B. The first class (nclass=1) are spectra very close to the Tdata's ones. These spectra are the ones we want to select. The second class corresponds to spectra that are very far from Tdata's spectra. The third class corresponds to spectra around the closure of Tdata in Ldata 2. If we apply GRSIR to the spectra from the first class, denoted *Ldata 2 selection PCA + GMM*, we can see that estimations are improved (see table 3.5). It confirms our idea that working on a reduced learning database leads to better estimations. On the contrary, this reduced database is not appropriate for the use of K-nn because it contains all the spectra that are close to the one observed, especially those that lead to the non unicity of solutions and deteriorate K-nn estimations. Moreover, the histograms of the SIR estimations allow to have an a priori about the range of variations of the parameters and a new learning database could be simulated to improve estimations. Figure 3.6.9 shows the histograms of the parameters estimated with GRSIR applied to Tdata with Ldata 2 (PCA + GMM). We can see that the range of variations of the estimated parameters correspond quite well to the reality. These observations also lead to the the conclusion that Ldata 1 is the most appropriate database to reverse Tdata.
In the following, we will denote *PCA + GMM methodology* the methodology we proposed to select an appropriate learning database for GRSIR.

	GRSIR				K-nn	
	Ldata 2 Selection		Ldata 2		Ldata 2 Selection	Ldata 2
	NRMSE	SIRC	NRMSE	SIRC	NRMSE	NRMSE
Prop. of water	**0.33**	**0.93**	0.63	0.88	0.53	0.86
Prop. of CO2	**0.23**	**0.98**	0.40	0.97	0.53	0.88
Prop. of dust	**0.16**	**0.99**	0.31	0.99	0.27	0.44
Grain size water	0.42	**0.86**	**0.41**	0.80	0.44	0.43
Grain size CO2	**0.20**	**0.98**	0.27	0.93	0.38	0.53

Table 3.4: NRMSE and SIRC criteria for Ldata 2 and Ldata 2 selection



A B

Figure 3.6.8: Selection of the Learning database. PCA + Gaussian mixture model (GMM)

	GRSIR		K-nn
	Selection PCA + GMM		Selection PCA + GMM
	NRMSE	SIRC	NRMSE
Prop. of water	**0.40**	0.90	0.87
Prop. of CO2	**0.30**	0.98	0.88
Prop. of dust	**0.17**	0.99	0.40
Grain size water	0.54	0.84	**0.35**
Grain size CO2	**0.22**	0.95	0.53

Table 3.5: Validation criteria for GRSIR and K-nn applied on Ldata 2 selection PCA + GMM

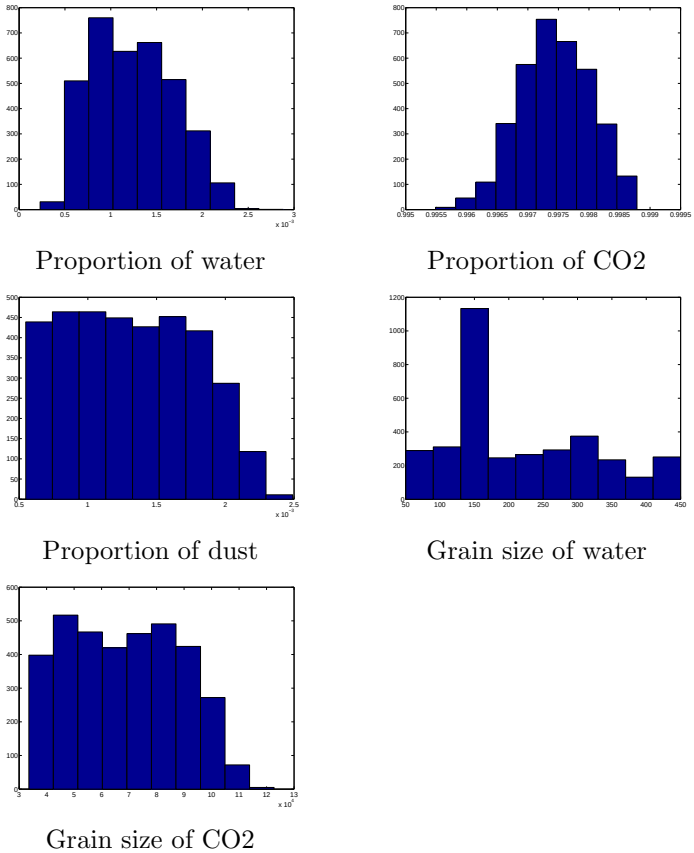


Figure 3.6.9: Histograms of the parameters estimations by GRSIR applied to Ldata 2 (PCA + GMM)

3.7 How to deal with dependent parameters?

Applying the Gaussian Regularized Sliced Inverse Regression to each parameter individually faces the problem that the sum of proportions of water, dust and CO2, will not be equal to 1 as it should be. Let us denote $\hat{y}_{dust}$ the estimated proportion of dust, $\hat{y}_{H2O}$ the estimated proportion of water and $\hat{y}_{CO2}$ the proportion of CO2 for one spectrum by Gaussian Regularized Sliced Inverse Regression. Let us denote $T = \hat{y}_{dust} + \hat{y}_{CO2} + \hat{y}_{H20}$ the sum of the estimated proportions. In this section we propose to compare the following methods to make the sum of proportions equal to 1:

- "N-GRSIR" (normalized SIR): we divide each estimated proportion $\hat{y}_{dust}$, $\hat{y}_{CO2}$ or $\hat{y}_{H20}$ by their sum T. Generally T is very close to one.
- "$1 - [CO2] - [dust]$" methodology: the proportions of CO2 and dust are estimated by SIR, then the proportion of water is deduced by $\hat{y}_{H20} = 1 - \hat{y}_{dust} - \hat{y}_{CO2}$. The problem of this method is that proportion of water can be negative.
- "$1 - [H20] - [dust]$" methodology: the proportions of water and dust are estimated by SIR, then the proportion of CO2 is deduced by $\hat{y}_{CO2} = 1 - \hat{y}_{dust} - \hat{y}_{H20}$.
- "C-GRSIR": because using the "$1 - [CO2] - [dust]$" methodology, proportions of water can be negative, we propose to keep this methodology when proportions of water are positive and to use "$1 - [H20] - [dust]$" methodology when they are negative.

Comparisons of these methods in terms of NRMSE are presented in table 3.6 for Ldata 1 and 3.7 for Ldata 2. "C-GRSIR" seems to be the best methodology to apply because it does not give negative estimations of proportions and it does not deteriorate estimations of the proportion of CO2 and dust and nor deteriorates significantly estimations of the proportion of water.

	Proportion of water	Proportion of CO2	Proportion of dust
GRSIR	0.29	**0.22**	**0.13**
K-nn	0.50	0.54	0.34
WK-nn	0.38	0.46	0.35
N-GRSIR	0.29	0.26	**0.13**
$1-[CO2]-[dust]$	**0.27**	**0.22**	**0.13**
$1-[H2O]-[dust]$	0.29	0.26	**0.13**
C-GRSIR	**0.27**	**0.22**	**0.13**

Table 3.6: Comparisons between the NRMSE's for the 7 proposed methods to estimate proportions. Learning database: Ldata 1.

	Proportion of water	Proportion of CO2	Proportion of dust
GRSIR	0.63	**0.40**	**0.31**
K-nn	0.86	0.88	0.44
WK-nn	**0.60**	0.68	0.41
N-GRSIR	0.63	0.52	**0.31**
$1-[CO2]-[dust]$	0.69	**0.40**	**0.31**
$1-[H2O]-[dust]$	0.63	0.52	**0.31**
C-GRSIR	0.68	**0.40**	**0.31**

Table 3.7: Comparisons between the NRMSE's for the 7 proposed methods to estimate proportions. Learning database: Ldata 2.

3.8 Final results

Finally, applying C-GRSIR to Ldata 1 and comparing the estimations of Tdata with K-nn shows that C-GRSIR gives in average, better estimations for the all set of parameters (see table 3.8).

Globally:

- for 93%, at least 1 parameter is stricly better estimated by C-GRSIR,
- for 79%, at least 2 parameters are stricly better estimated by C-GRSIR,
- for 56%, at least 3 parameters are stricly better estimated by C-GRSIR,
- for 42%, at least 4 parameters are stricly better estimated by C-GRSIR,
- for 20%, all parameters are stricly better estimated by C-GRSIR.

On the other hand:

- for 77%, at least 1 parameter is stricly better estimated by K-nn,
- for 52%, at least 2 parameters are stricly better estimated by K-nn,

- for 31%, at least 3 parameters are stricly better estimated by K-nn,
- for 17%, at least 4 parameters are stricly better estimated by K-nn,
- for 4%, all parameters are stricly better estimated by K-nn.

	K-nn	WK-nn	SIR	
	NRMSE	NRMSE	NRMSE	SIRC
Proportion of water	0.50	0.38	**0.27**	0.92
Proportion of CO2	0.54	0.46	**0.22**	0.99
Proportion of dust	0.34	0.35	**0.13**	0.99
Grain size of water	0.39	0.45	**0.37**	0.92
Grain size of CO2	0.35	0.46	**0.19**	0.98

Table 3.8: Final results: K-nn and WK-nn with Ldata 1. Tikhonov C-GRSIR with Ldata 1.

Chapter 4

Application to real data

In this chapter, we present the inversion of images acquired on Mars during orbit 41, 30, 61 and 103. K-nn, WK-nn and C-GRSIR inversions are compared.
A few points are first discussed before the final inversions:

- Because no ground data is available, validation is difficult. In section 4.1, we discuss a possible way to validate results.
- The noise in real images can be really important for some of the wavelengths and can lead to biases in the estimations. We show in section 4.2 that it can be judicious to remove some of the wavelengths before the inversion. If not, the regularization parameter should be increased.
- When the spectra are very different from the one simulated in the learning database, we can wonder if it is appropriate to reverse these spectra and we propose in section 4.3 a way to select the "invertible" spectra from the real image.

Finally, the GRSIR methodology is summarized in section 4.4 and results are given in section 4.5.

4.1 How to validate results for a real data?

In the previous chapter, we validated our results thanks to synthetic data (spectra generated by a model) with the help of two validation criteria. In the case of a real hyperspectral from Mars, it becomes impossible to validate directly the results given by C-GRSIR because no ground images are available. We thought it could be a good idea to simulate spectra for parameters values estimated at each pixel by C-GRSIR and to compare them with the observed spectra. It amounts to reconstituting an hyperspectral cube with C-GRSIR estimations. We have seen previously that two spectra can be very close in terms of RMSE even if the associated parameters are very different. This shows that methods cannot be compared according to the RMSE between observed spectra and reconstituted spectra. However, we can hope that if parameters are close then spectra are close and we hope that reconstituting an hyperspectral cube can help to evaluate the quality of our estimations.

We have seen previously that C-GRSIR applied to Ldata 1 in order to estimate Tdata gave the best parameters estimations for most of the spectra. If we now simulate an hyperspectral cube with C-GRSIR parameters estimations, we can see that generally, reconstituted spectra are indeed very close from the ones observed in Tdata. Such an example is given in figure 4.1.1. The mean square errors between test spectra and reconstituted spectra in average for all pixels are given for C-GRSIR and K-nn methodology in table 4.1. Of course, in that case, results are better for K-nn methodology because this method precisely searches for the best match between observed spectra and simulated spectra in terms of Mean Square Errors.

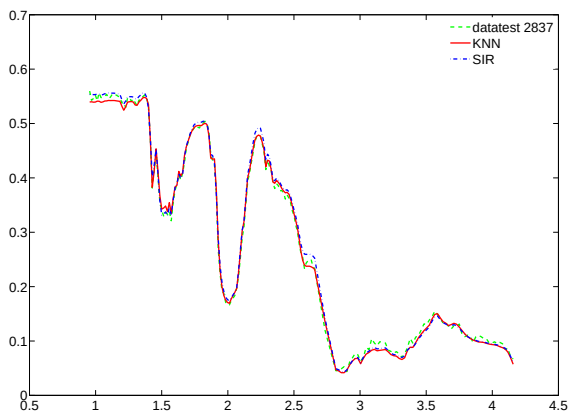


Figure 4.1.1: Analyzing the K-nn results for one spectrum of Tdata. Ldata 2. Horizontally: Wavelengths, Vertically: Reflectance.

	SIR	KNN
Mean square errors	0.0084	0.0053

Table 4.1: Mean square errors (cf equation 3.2.2) between spectra from Tdata and reconstituted spectra for SIR and K-nn in average for all pixels. Learning database: Ldata 1.

4.2 Masking some of the wavelengths?

When we first applied C-GRSIR to image 41 with Ldata 1, estimations were very different from the ones estimated by K-nn especially for the grain size of CO2. In fact, we have seen that with real observations, the reflectance of some limited spectral intervals cannot be reproduced systematically well by the model. The origins of such discrepancies can be numerous:

- instrument calibration problem,
- deficiency of the atmospheric or thermal corrections,

- contribution of an unknown compound not taken into account,
- flaws in the physics of light reflection by complex media.

All these points can be improved with an increase of knowledge but currently lead to biases in SIR estimations. For example systematic discrepancies (including noise) are greater than 10% for 26% of the wavelengths and greater than 20% for 13% of the wavelengths whereas the expected noise for image 41 is of the order of 1-2%. In order to reduce the bias due to the systematic misfit of some wavelengths, we decided to work only with the wavenlengths for which the misfit is estimated to be less than 10% (selection 10) or 20% (selection 20) on average, based on a first analysis of image 41 with the K-nn method. We can wonder if working on a reduced number of wavelengths is going to strongly deteriorate estimations, so we tested C-GRSIR on Tdata with selection 10 and selection 20 and compared the resulting validation criteria with those obtained on 184 wavelengths. Results (table 4.2) show that it does not deteroriate significantly results and that working on a selection of wavelengths is a good compromise to get rid of discrepancies. An other way to proceed is to consider that if the inversion by C-GRSIR gives many estimations close to the minimum or maximum value, it is because the noise on the observed data is greater or smaller than the one introduced in the SIR methodology and consequently the choosen regularization parameter is not appropriate and should be increased or reduced. We show in appendix E, the evolution of image 41's inversion according to the choosen regularization parameter for the all set of parameters (proportion of water has to be interpreted with care because in a way, it is not estimated by GRSIR but just deduced from the others proportions). We can see that if the regularization parameter is too small or on the contrary too high, then a consequent number of pixels are estimated to the minimum or maximum value. But there exists a range of values for which the inversion give a "smoother" histogram. Generally, the regularization parameter choosen by the minimization of the NRMSE criterion belongs to this range of values. But if not, we propose to increase or decrease the regularization parameter in order to be in this range of values. In practice, we have seen that this way of proceeding gives similar results that the use of a mask on wavelengths. We will prefer this methodology because it does not require the use of K-nn on the contrary to selecting some of the wavelengths.

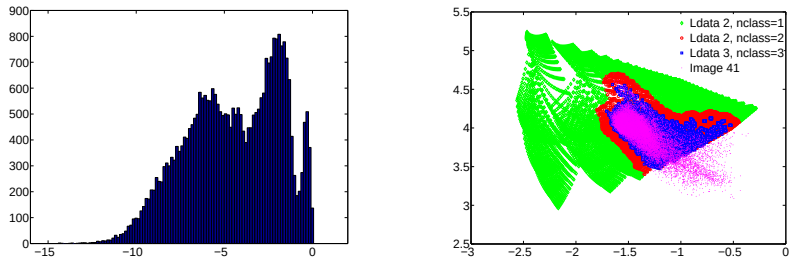
Parameters	No selection		Selection 20		Selection 10	
	NRMSE	SIRC	NRMSE	SIRC	NRMSE	SIRC
Proportion of water	0.40	**0.90**	0.40	0.89	**0.37**	0.89
Proportion of CO2	**0.30**	**0.98**	0.31	**0.98**	0.31	**0.98**
Proportion of dust	0.17	**0.99**	0.19	**0.99**	**0.16**	**0.99**
Grain size of water	**0.54**	0.84	**0.54**	0.84	**0.54**	**0.85**
Grain size of CO2	**0.22**	**0.95**	0.24	**0.95**	0.26	0.94

Table 4.2: Comparisons between different sampling strategies to select valid wavelengths for the study. Methodology: Tikhonov GRSIR. Estimation of the test data parameters by Ldata 1.

4.3 Selection of the learning database

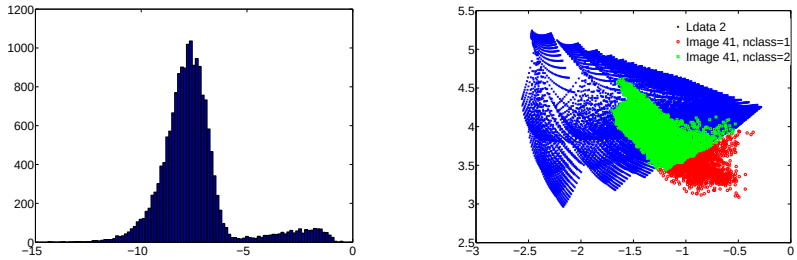
We have seen in section 1.3.3 that some of the spectra in real images from Mars cannot be reversed because the physical model is not relevant for them. These spectra should then be removed from the inversion. We propose to select them by *PCA + GMM methodology*.
We present in figure 4.3.2 the *PCA + GMM methodology* applied to Ldata 2 and spectra observed from orbit 41 in order to select the most relevant spectra from Ldata 2. On the opposite, we present figure 4.3.3 the selection of the *invertible* spectra from image 41. The distances between spectra from image 41 and their nearest neighbors in Ldata 2 have been calculated. The histogram presented figure 4.3.3 A allows to distinguish two classes: spectra from image 41 that are outside Ldata 2, and spectra that are inside (see figure 4.3.3 B). For inversion, we will only select spectra that are *invertible* (nclass 2).

Each time an hyperspectral image has to be reversed, we propose to select the appropriate learning database applying PCA + GMM methodology, but also to select the spectra from the image that are considered invertible. The latter will also be selected by PCA + GMM methodology. Then the inversion of the image should be processed using GRSIR.



A B

Figure 4.3.2: Selection of the spectra in Ldata 2 for the inversion of the hyperspectral image 41. On the left: histogram of the logarithm of the distances from each point from Ldata 2 to its nearest neighbor in image 41. On the right: Selections of the classes by GMM.



A B

Figure 4.3.3: Selection of the invertible spectra in image 41. On the left: histogram of the logarithm of the distances from each point from the image 41 to its nearest neighbor in Ldata 2. On the right: Selections of the classes by GMM.

4.4 Final GRSIR methodology

Finally, in this section, we present a diagram (figure 4.4.4) summarizing the entire C-GRSIR methodology used for the inversion of any observed hyperspectral image. This way of proceeding will be used for the inversion of the images acquired during orbit 41, 30, 61 and 103 in the next section.

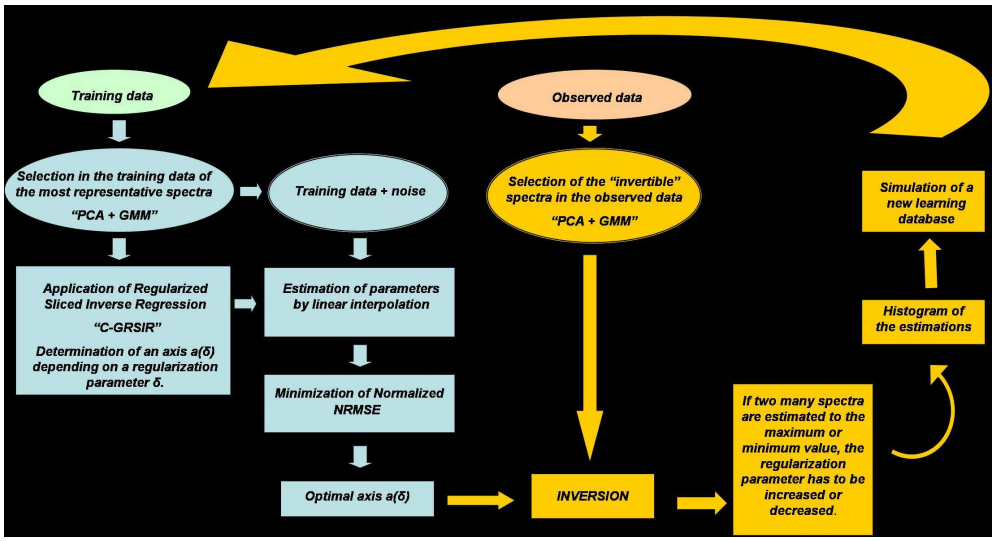


Figure 4.4.4: Final Regularized Sliced Inverse Regression methodology proposed to retrieve the physical parameters from an hyperspectral image

4.5 Results

In this section we present the first inversions obtained for image 41, 30, 61 and 103 with K-nn, WK-nn and C-GRSIR. K-nn and WK-nn inversions have been realized with Ldata 2 whereas C-GRSIR inversion has been realized with a selection of spectra from Ldata 2 determined by PCA + GMM methodology. For some of the parameters, the regularization parameter has been increased deteriorating the SIRC and NRMSE criteria but allowing the inversion. Results show that most of the time, C-GRSIR gives a very smooth mapping for the all set of parameters whereas with K-nn and WK-nn estimations can differ much more between two neighbor pixels. Moreover, it can happen in K-nn and WK-nn that only very few values of the learning database are retained which give the impression of a segmentation map more than an estimation map. For example, for the proportion of dust in image 103 only four values has been selected by K-nn and WK-nn. In most of the inversions, K-nn, WK-nn and C-GRSIR give inversions that are not in contradiction but estimations are slightly different. For example, in the inversion of image 103, proportion of CO2 is estimated in a range of 0.998-0.9995 with C-GRSIR whereas it varies in the range 0.996-0.9998 with K-nn. An interesting remark is about the estimation of the grain size of CO2 in images 61 and 103. These images represent nearly the same portion of surface of Mars and consequently, estimations of the grain size of CO2 should be approximately the same which is the case for C-GRSIR but not for K-nn and WK-nn which give much greater values for image 103. As we told before, it is difficult to tell if C-GRSIR gives better estimations than K-nn and WK-nn for real images. It would then be necessary to develop a methodology to associate uncertainties to estimations. This can be realized empirically, associating the experimental uncertainties deduced from the simulation of a test data, but it supposes that the test data is representative of the observed data and more especially that the introduced noise has been well evaluated. If not, uncertainties will be under or overestimated. In Sliced Inverse Regression, The SIRC and NRMSE criteria associated to parameters for each image give a first idea of the quality of the estimations and allow to deduce if one image seems to be better estimated than an other. When SIRC is smaller than 0.85 or NRMSE greater than 0.40, inversions are generally not smooth and doubtful. Comparisons between GRSIR and K-NN have to be discussed with care for the proportion of water because in GRSIR, proportion of water is deduced from the proportions of dust and CO2. That is also why in results, no SIRC is available for the proportion of water.

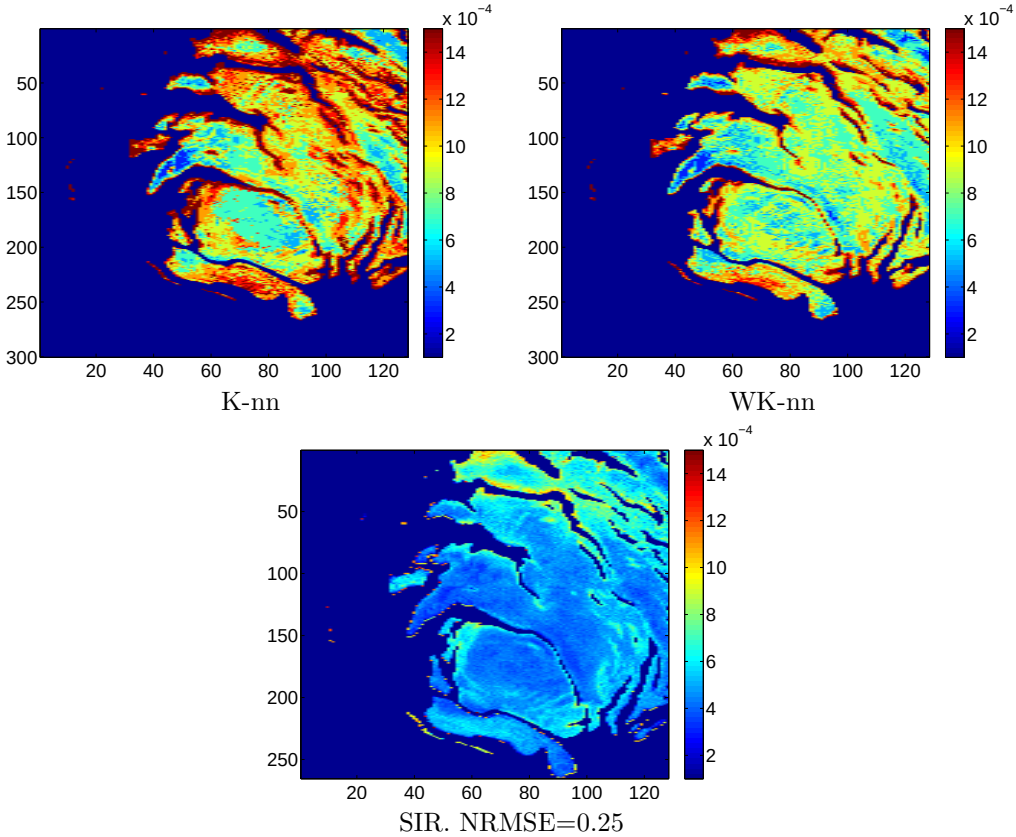


Figure 4.5.5: Studied image: during orbit 41. Proportion of water.

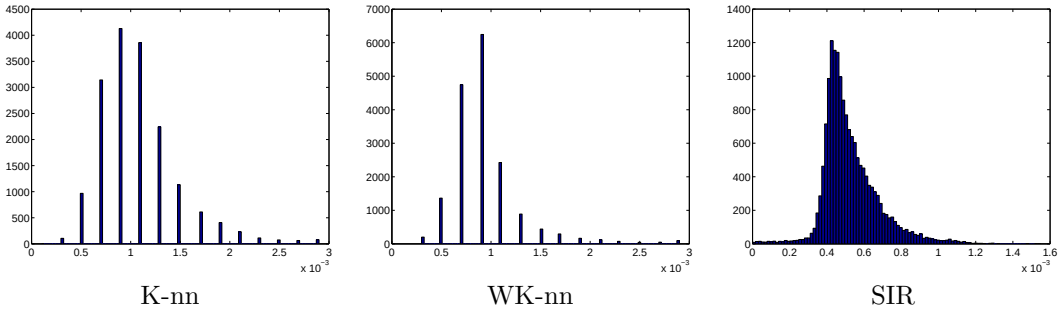


Figure 4.5.6: Studied image: during orbit 41. Histogram of the proportion of water.

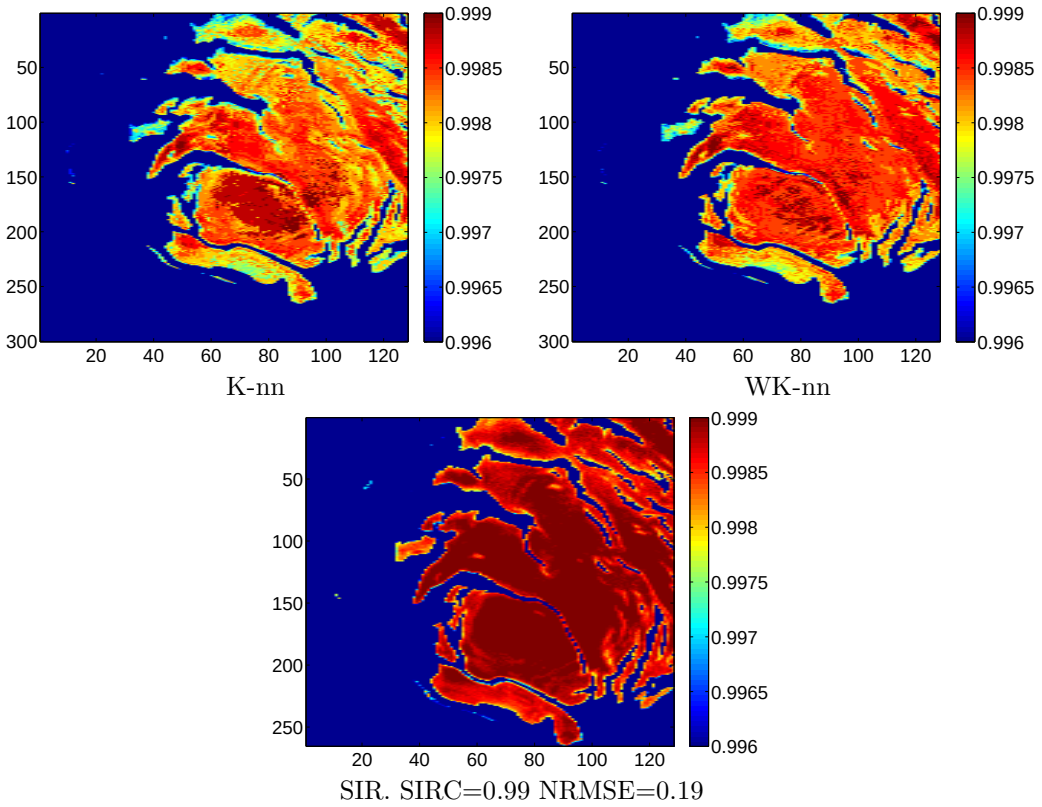


Figure 4.5.7: Studied image: during orbit 41. Proportion of CO2.

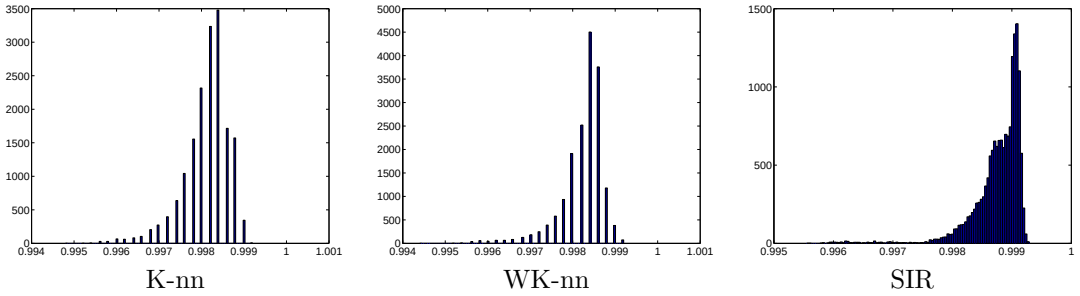
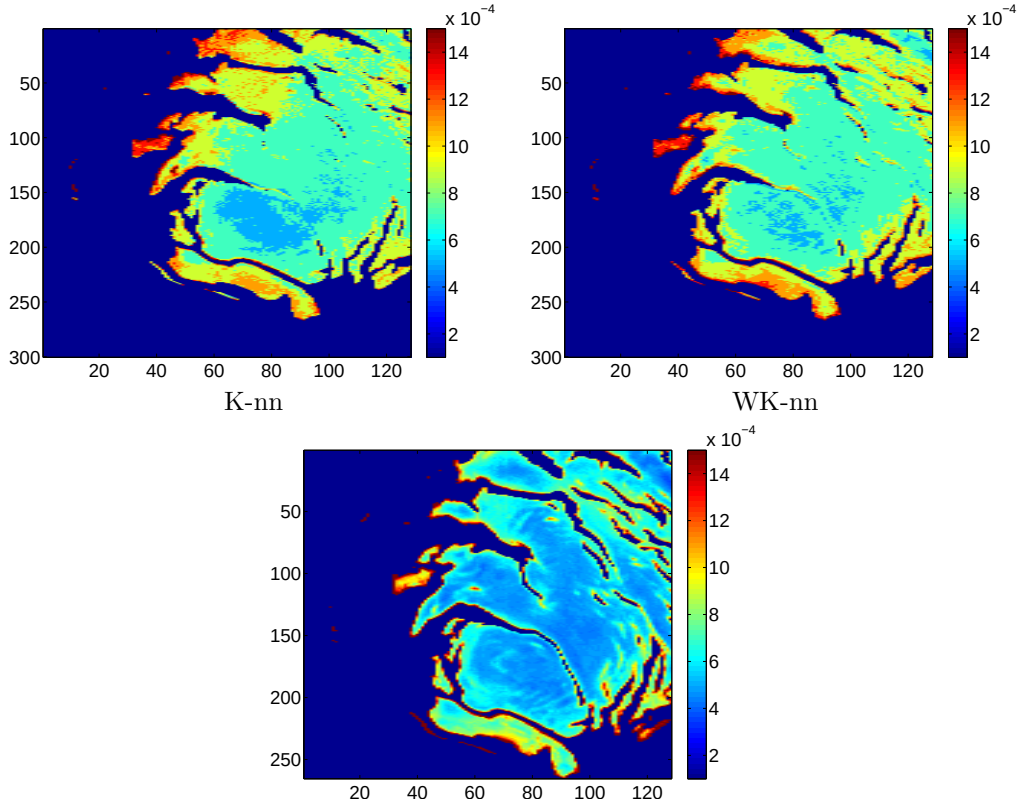


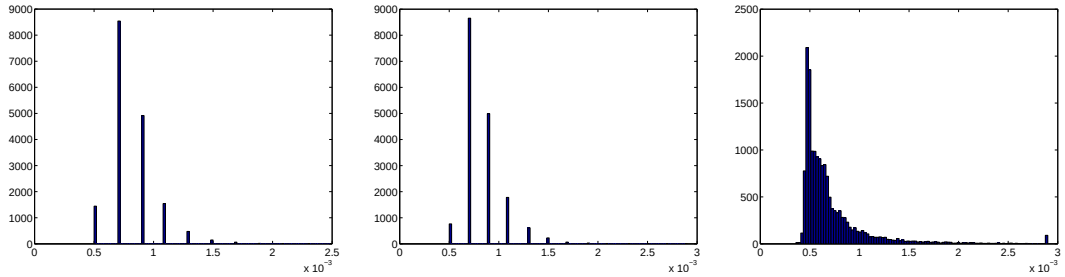
Figure 4.5.8: Studied image: during orbit 41. Histogram of the proportion of CO2.



K-nn WK-nn

SIR. SIRC=0.99 NRMSE=0.09

Figure 4.5.9: Studied image: during orbit 41. Proportion of dust.



K-nn WK-nn SIR

Figure 4.5.10: Studied image: during orbit 41. Histogram of the proportion of dust.

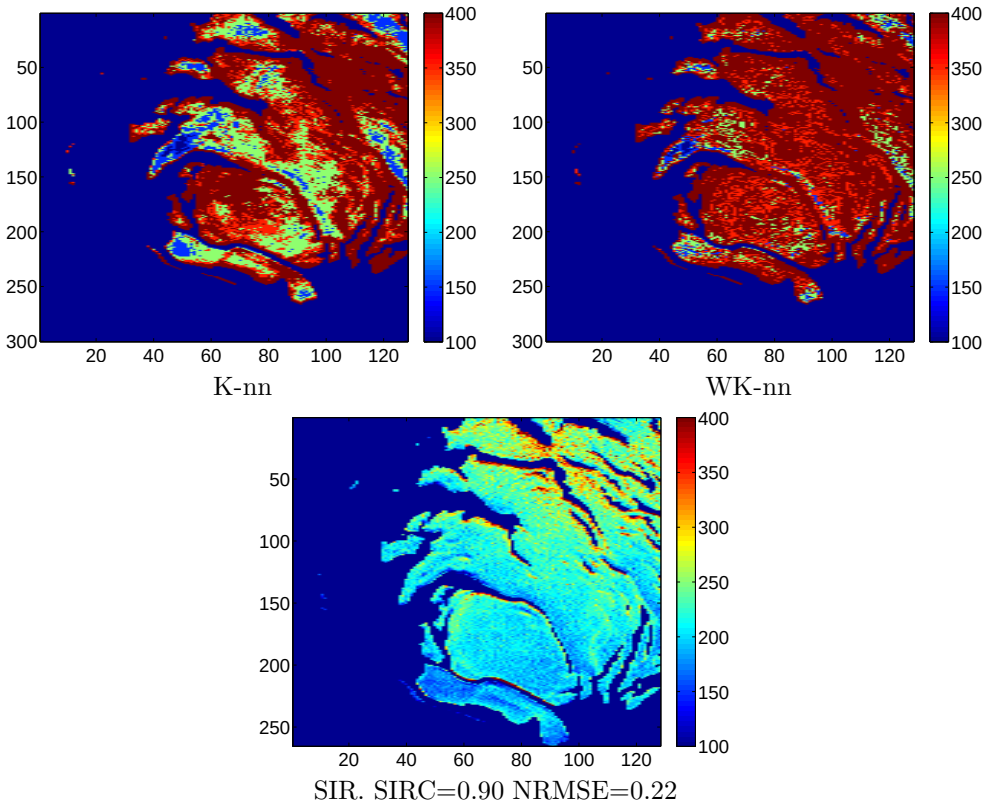


Figure 4.5.11: Studied image: during orbit 41. Grain size of water.

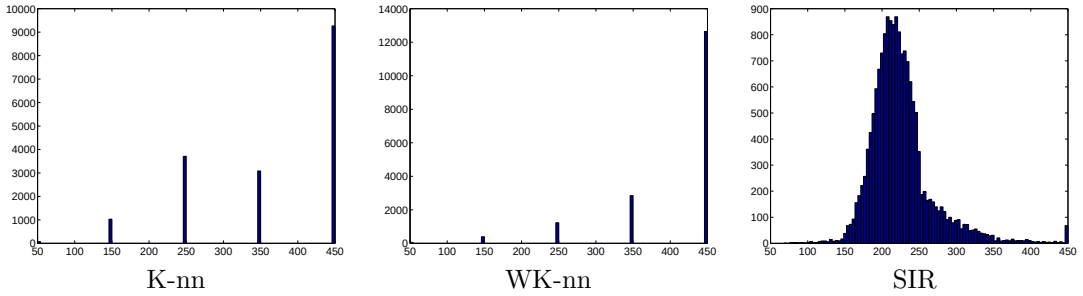


Figure 4.5.12: Studied image: during orbit 41. Histogram of the grain size of water.

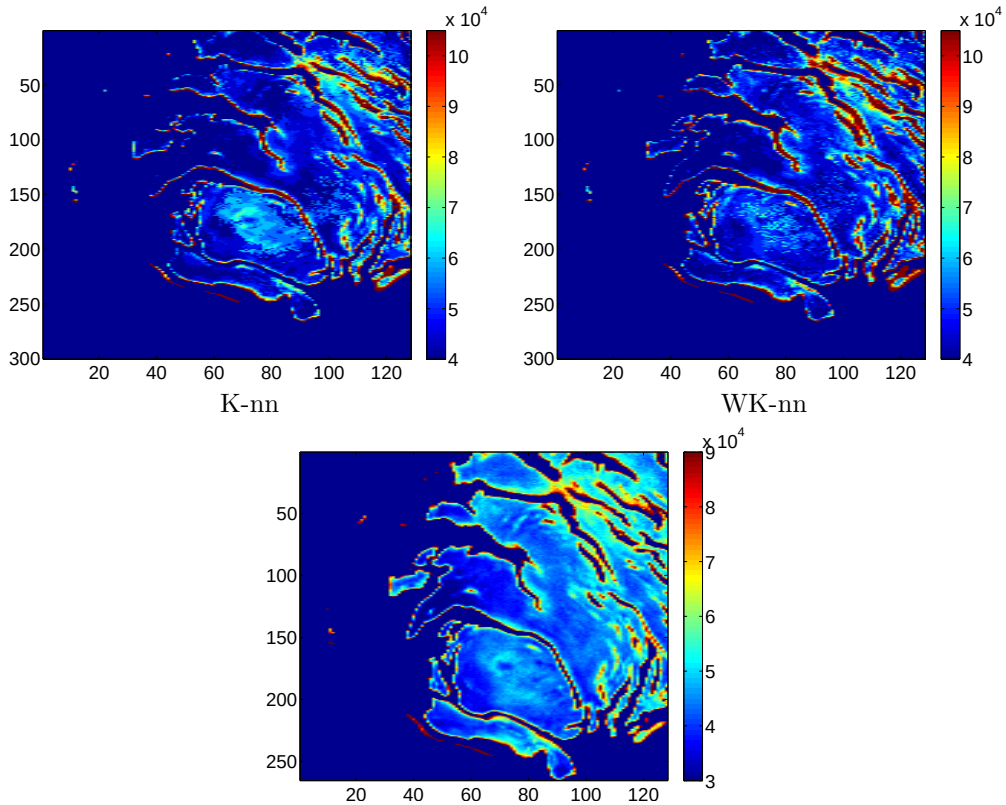


Figure 4.5.13: Studied image: during orbit 41. Grain size of CO2.

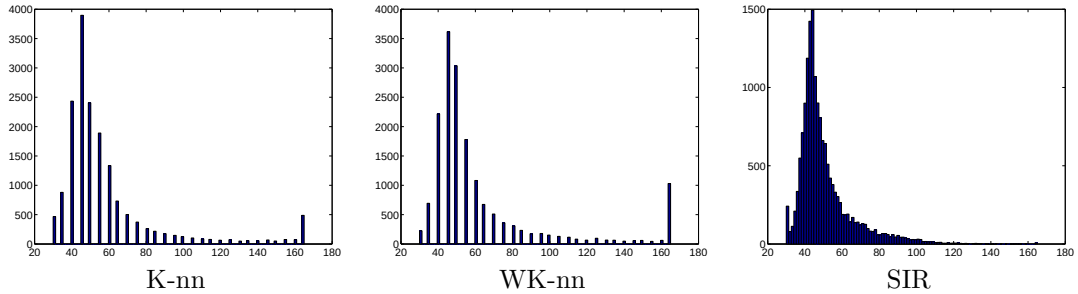


Figure 4.5.14: Studied image: during orbit 41.

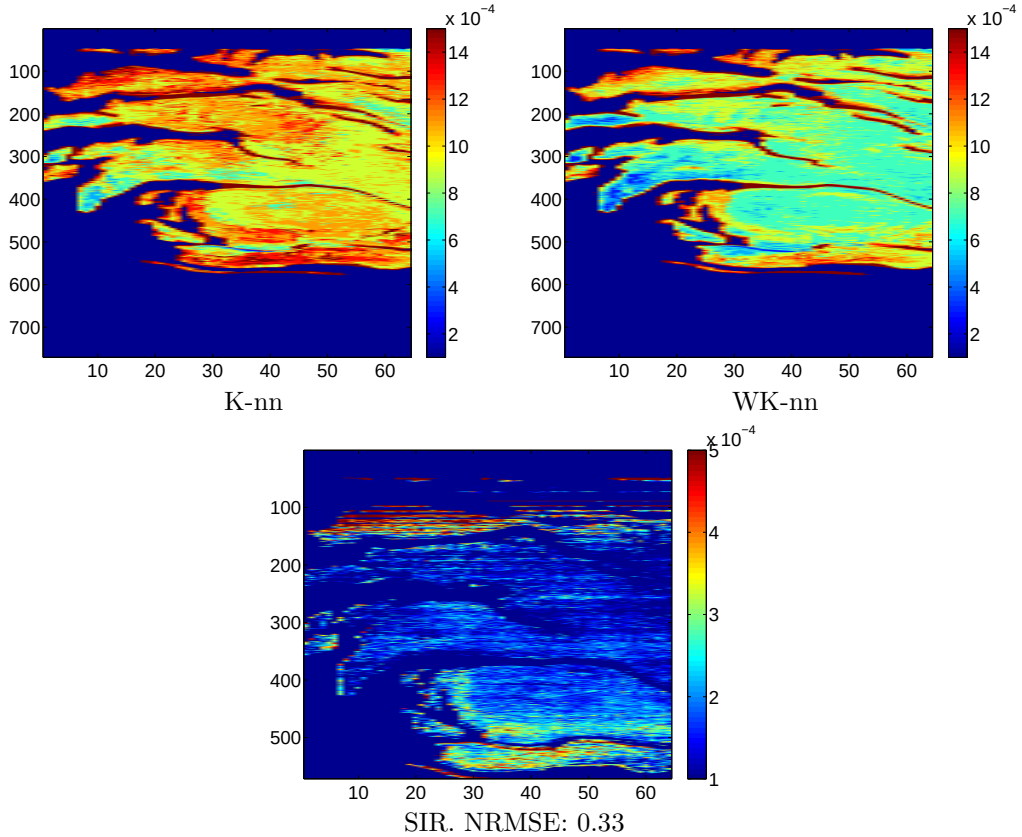


Figure 4.5.15: Studied image: during orbit 30. Proportion of water.

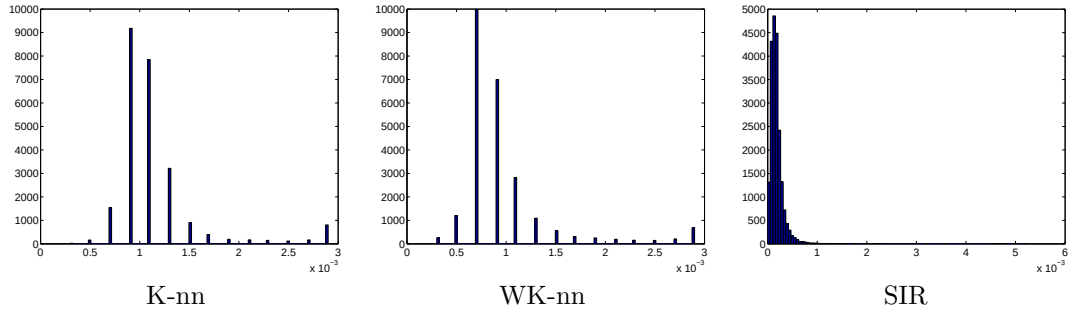


Figure 4.5.16: Studied image: during orbit 30. Histogram of the proportion of water.

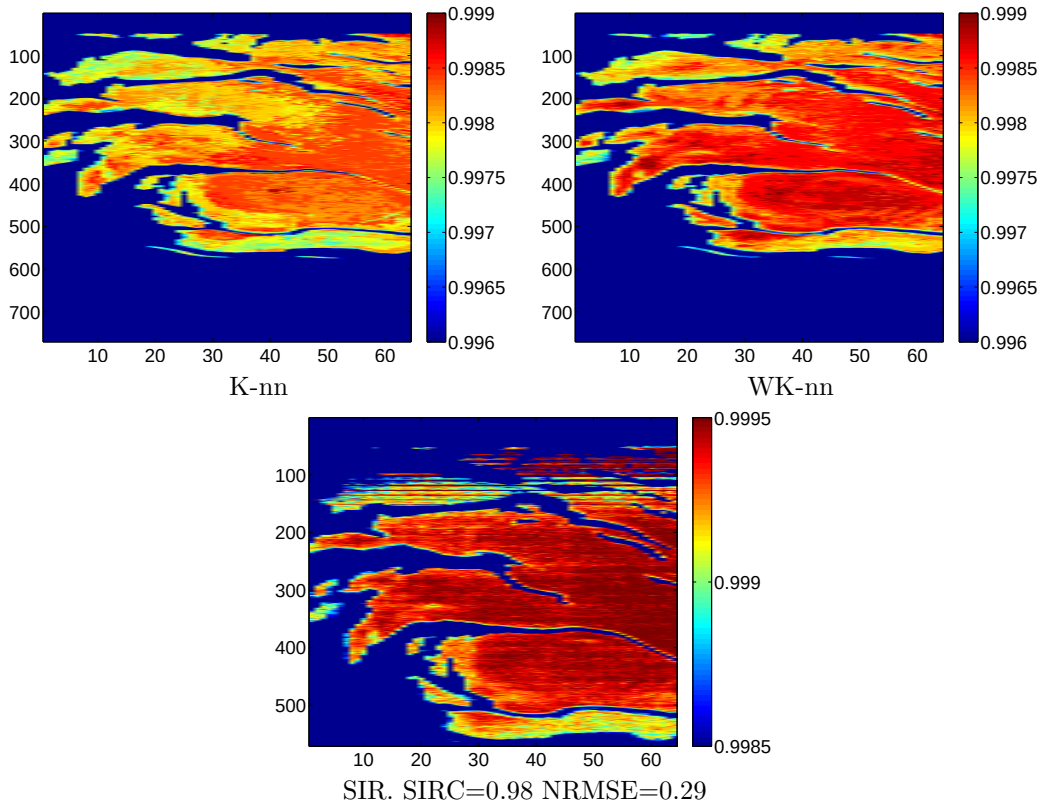


Figure 4.5.17: Studied image: during orbit 30. Proportion of CO2.

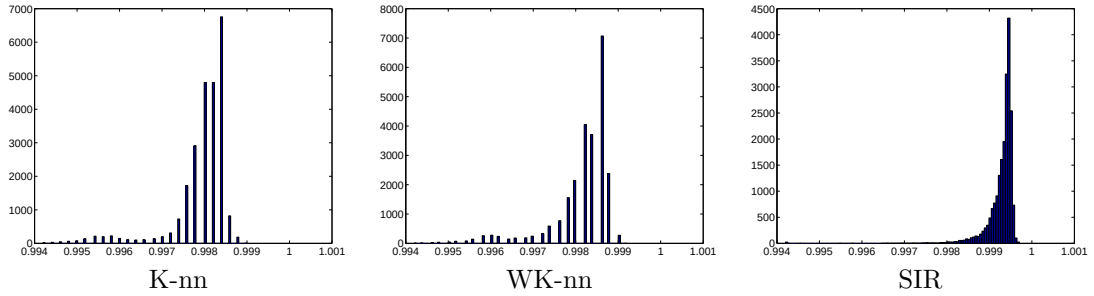


Figure 4.5.18: Studied image: during orbit 30. Histogram of the proportion of CO2.

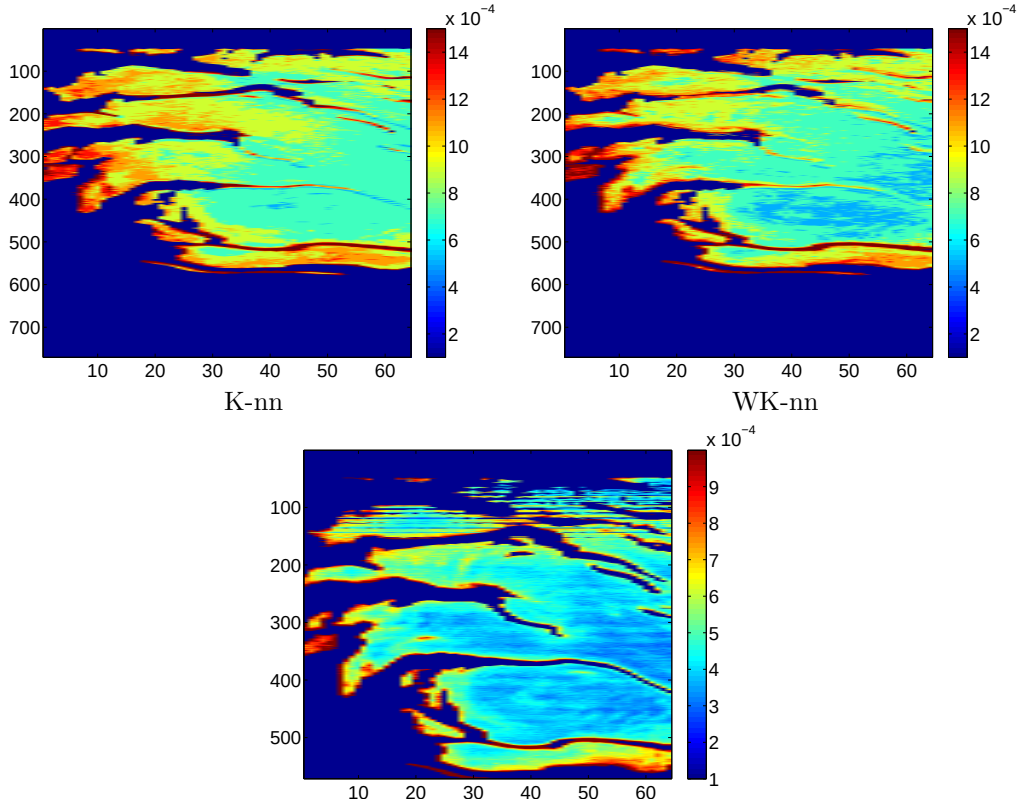


Figure 4.5.19: Studied image: during orbit 30. Proportion of dust.

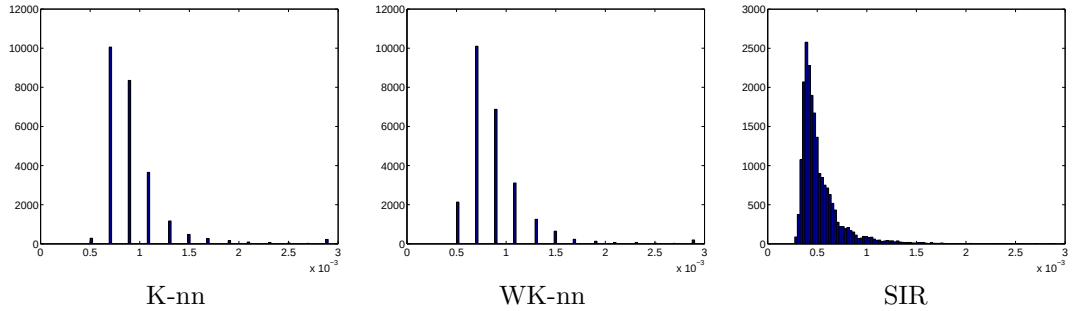


Figure 4.5.20: Studied image: during orbit 30. Histogram of the proportion of dust.

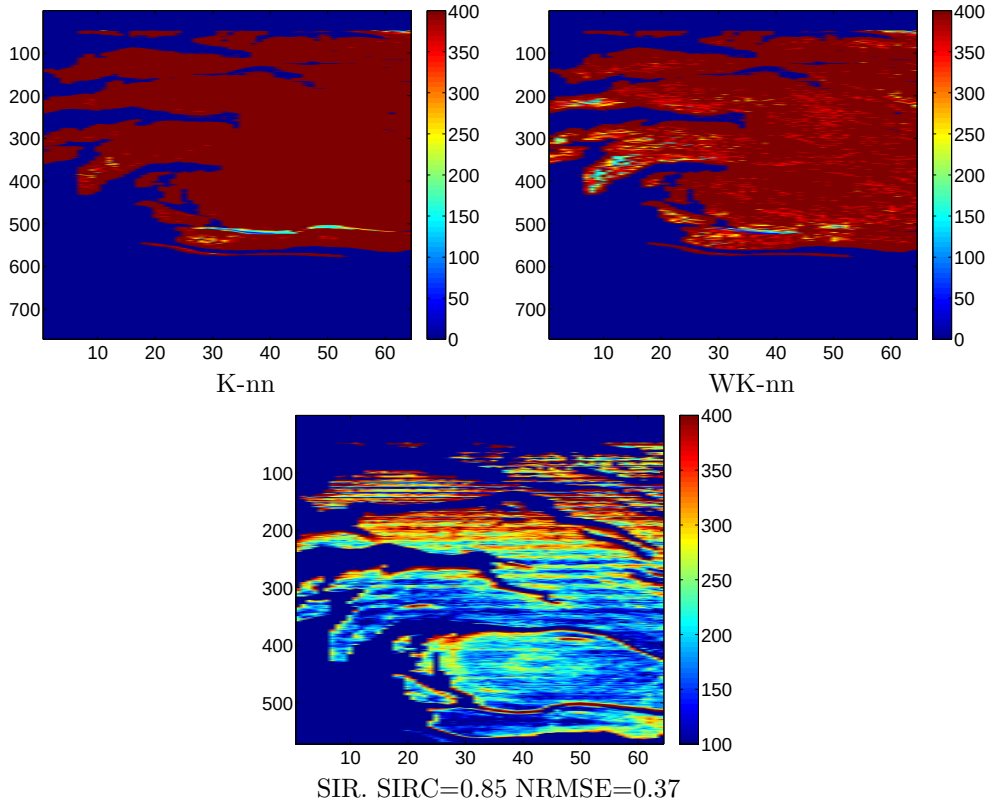


Figure 4.5.21: Studied image: during orbit 30. Grain size of water.

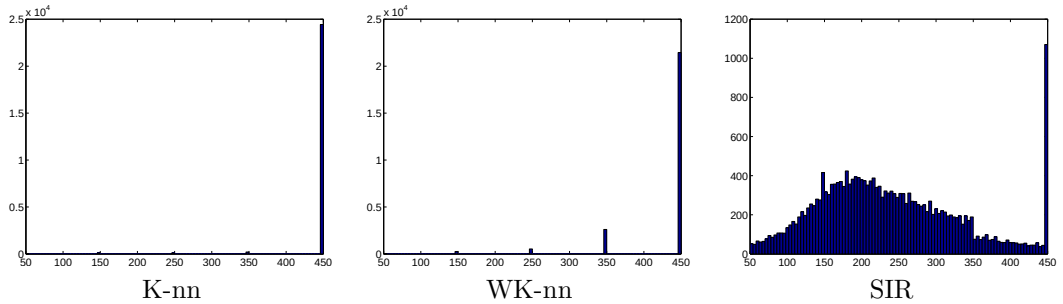


Figure 4.5.22: Studied image: during orbit 30. Histogram of the grain size of water.

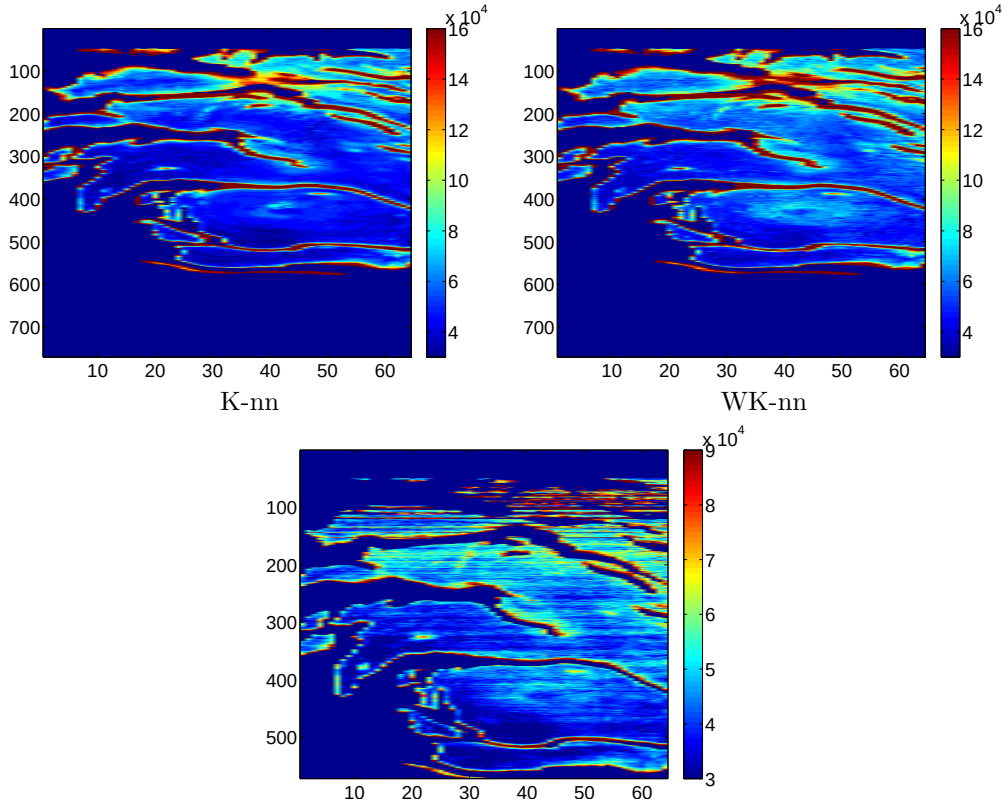


Figure 4.5.23: Studied image: during orbit 30. Grain size of CO2.

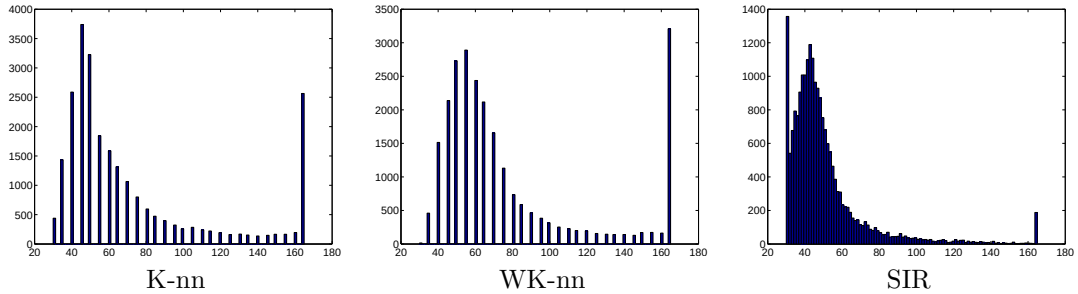


Figure 4.5.24: Studied image: during orbit 30.

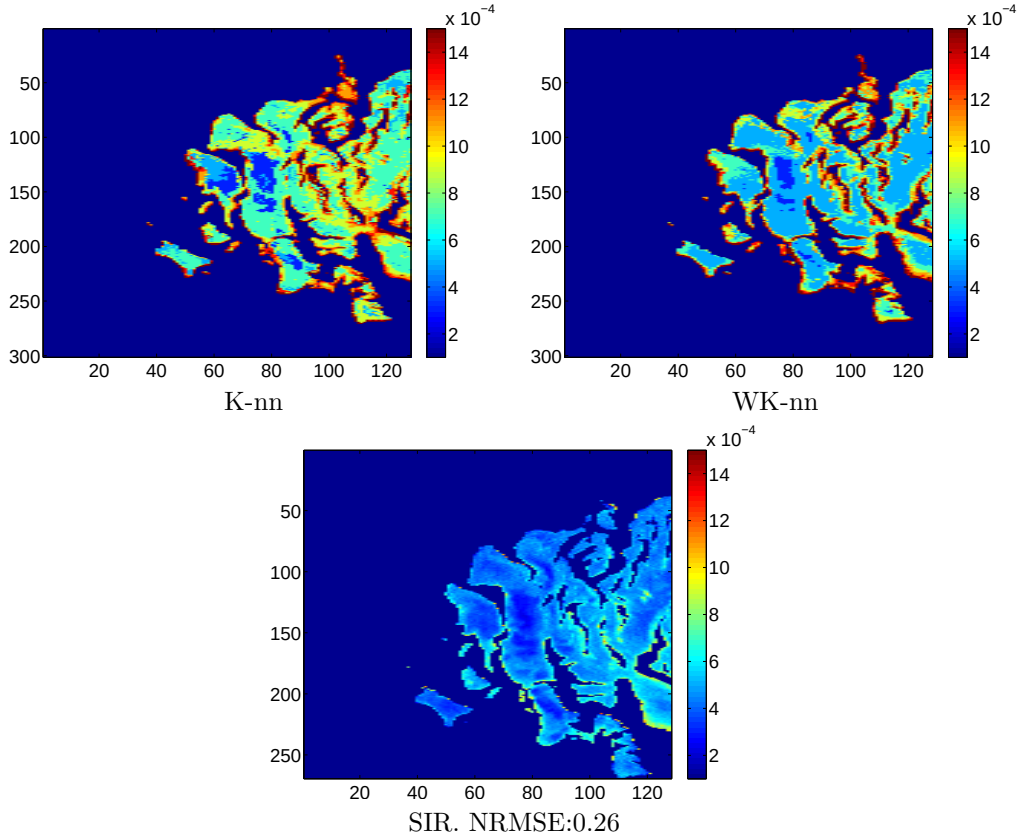


Figure 4.5.25: Studied image: during orbit 61. Proportion of water.

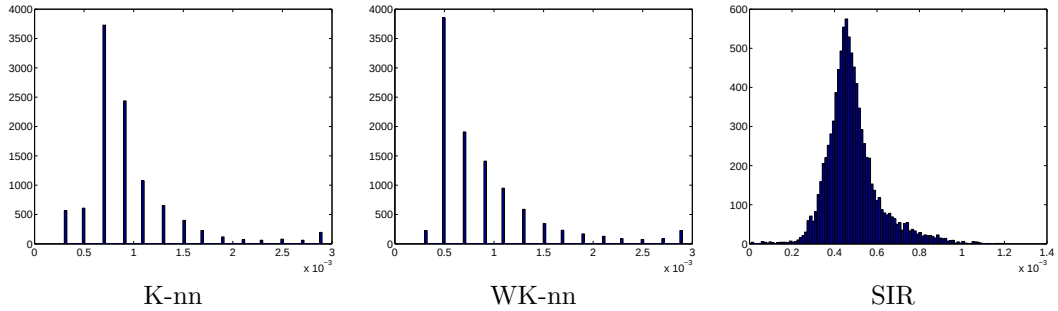


Figure 4.5.26: Studied image: during orbit 61. Histogram of the proportion of water.

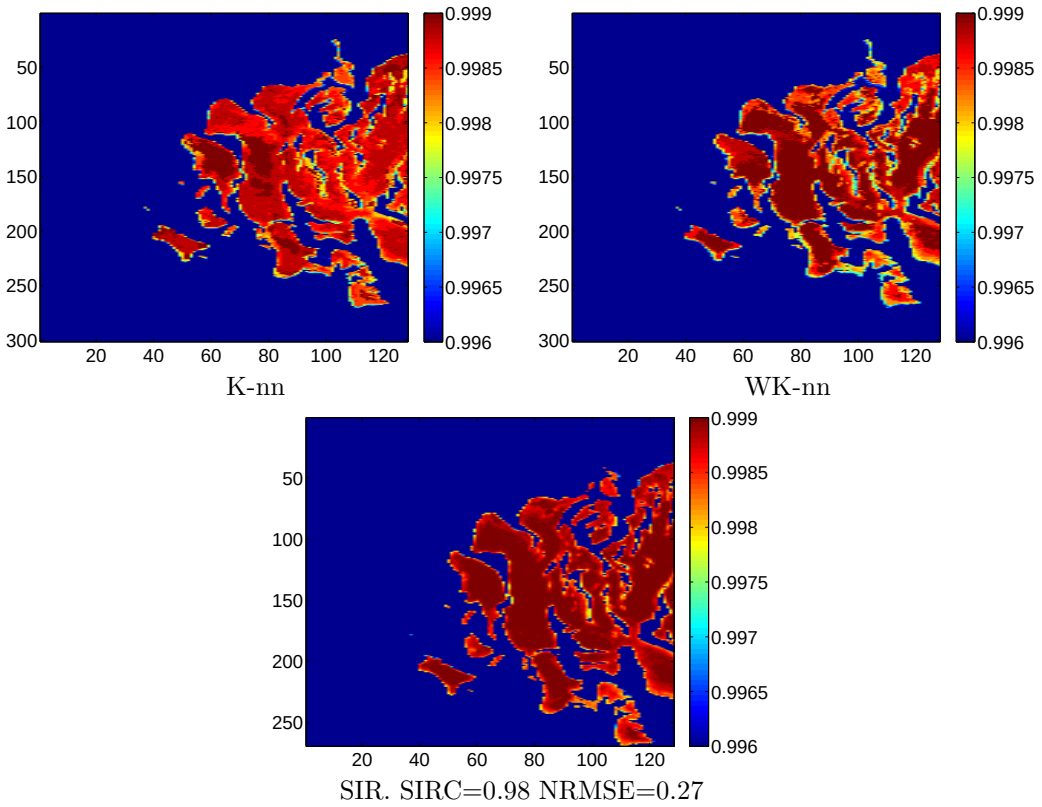


Figure 4.5.27: Studied image: during orbit 61. Proportion of CO2.

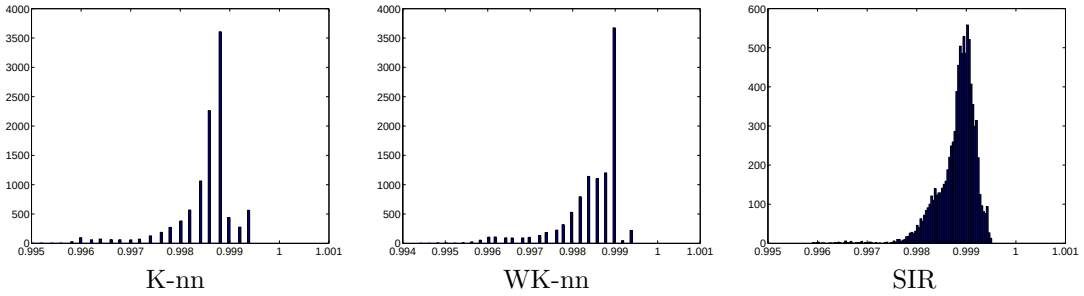
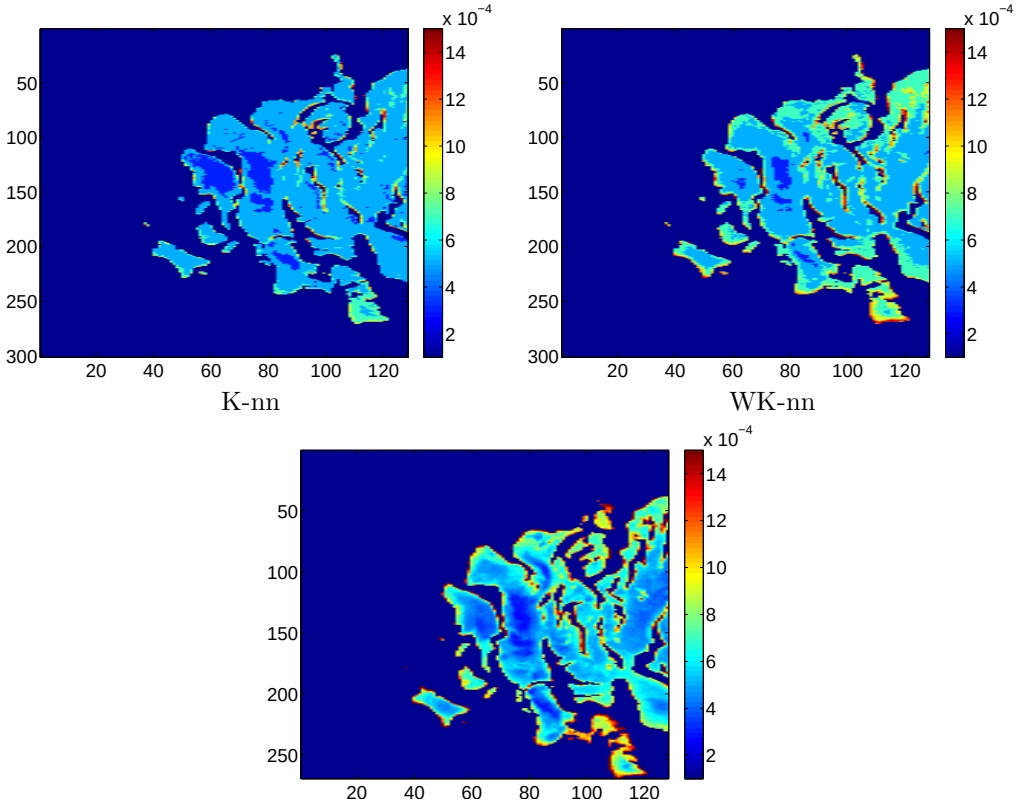


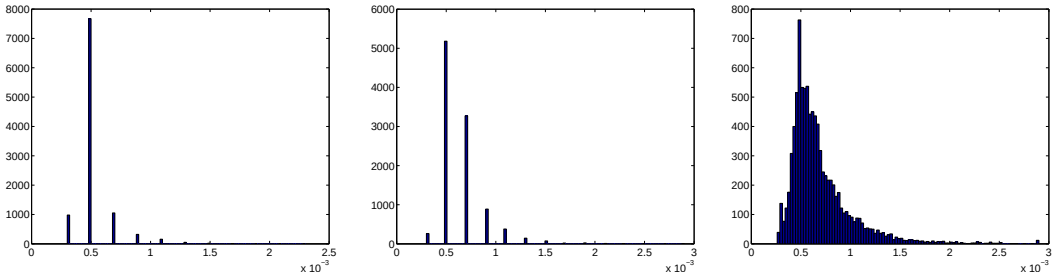
Figure 4.5.28: Studied image: during orbit 61. Histogram of the proportion of CO2.



K-nn WK-nn

SIR. SIRC=0.99 NRMSE=0.14

Figure 4.5.29: Studied image: during orbit 61. Proportion of dust.



K-nn WK-nn SIR

Figure 4.5.30: Studied image: during orbit 61. Histogram of the proportion of dust.

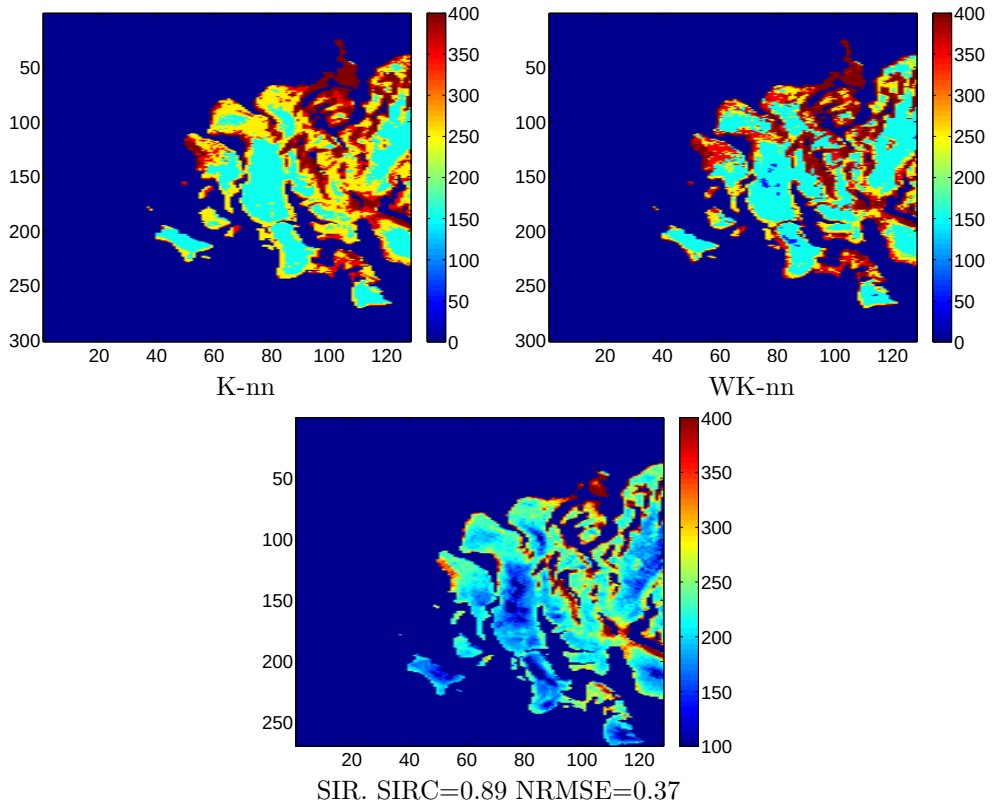


Figure 4.5.31: Studied image: during orbit 61. Grain size of water.

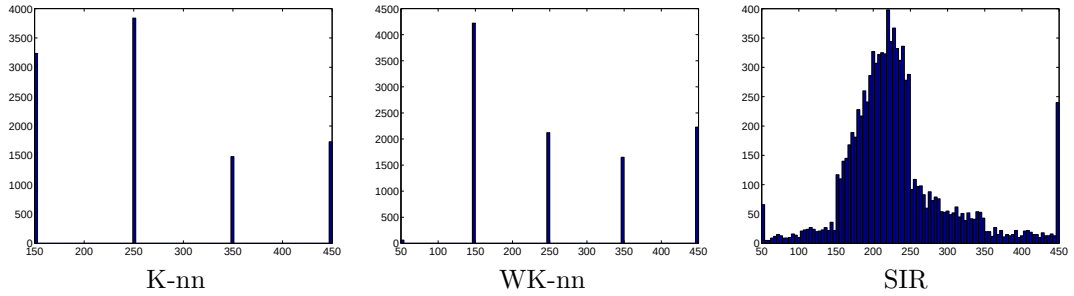


Figure 4.5.32: Studied image: during orbit 61. Histogram of the grain size of water.

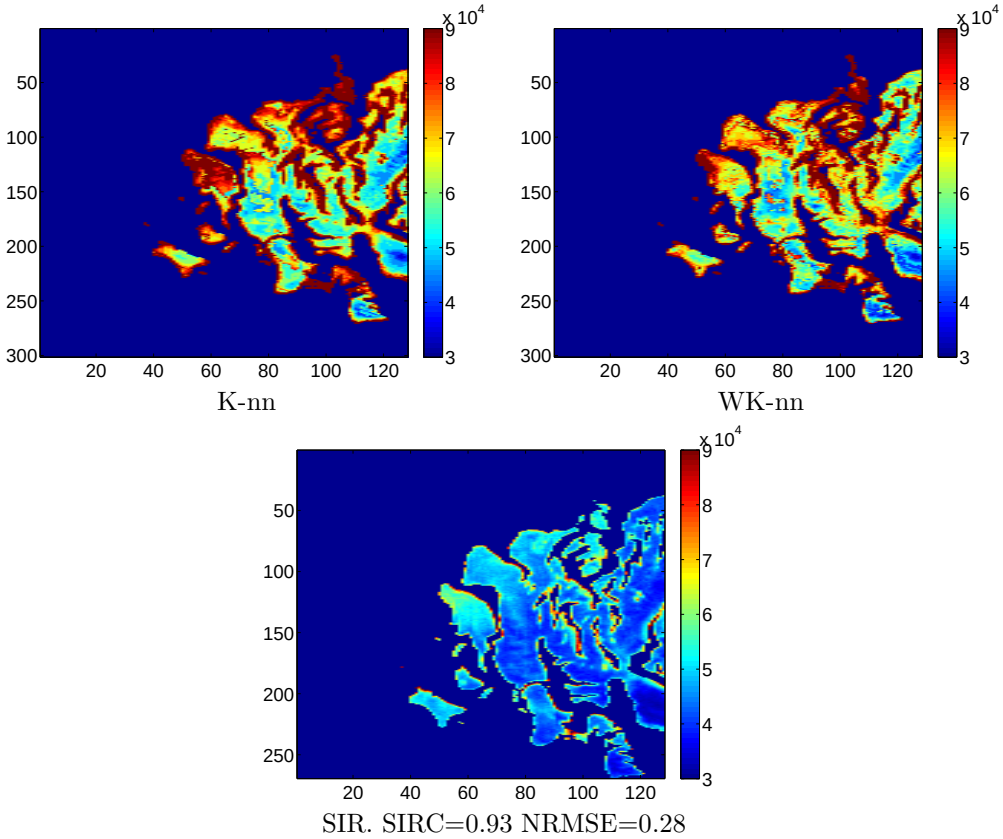


Figure 4.5.33: Studied image: during orbit 61. Grain size of CO2.

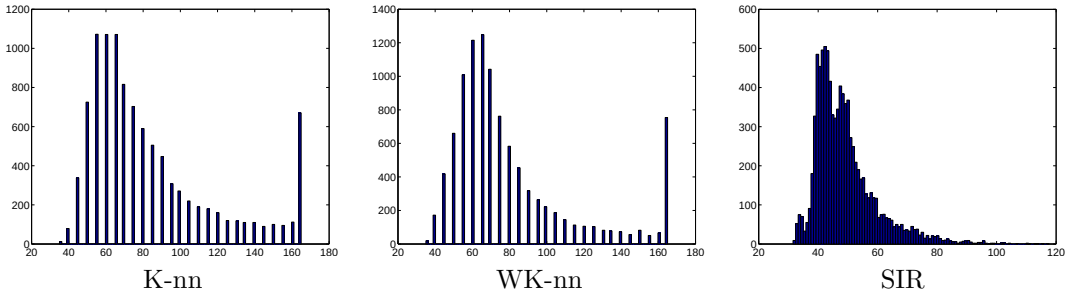


Figure 4.5.34: Studied image: during orbit 61.

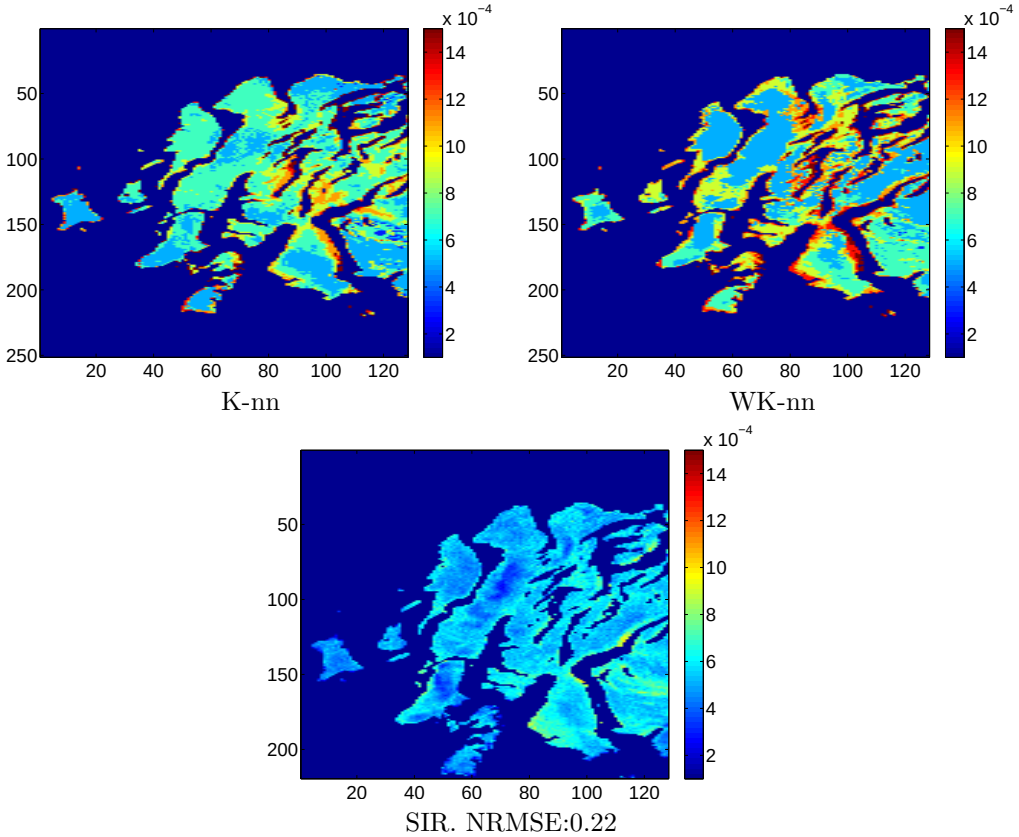


Figure 4.5.35: Studied image: during orbit 103. Proportion of water.

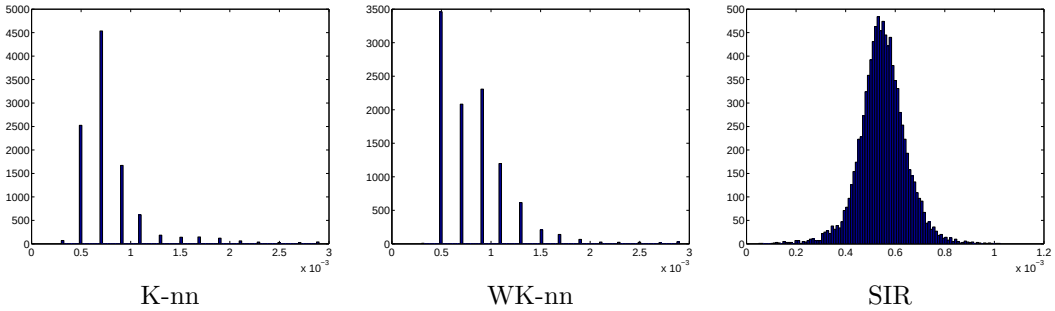


Figure 4.5.36: Studied image: during orbit 103. Histogram of the proportion of water.

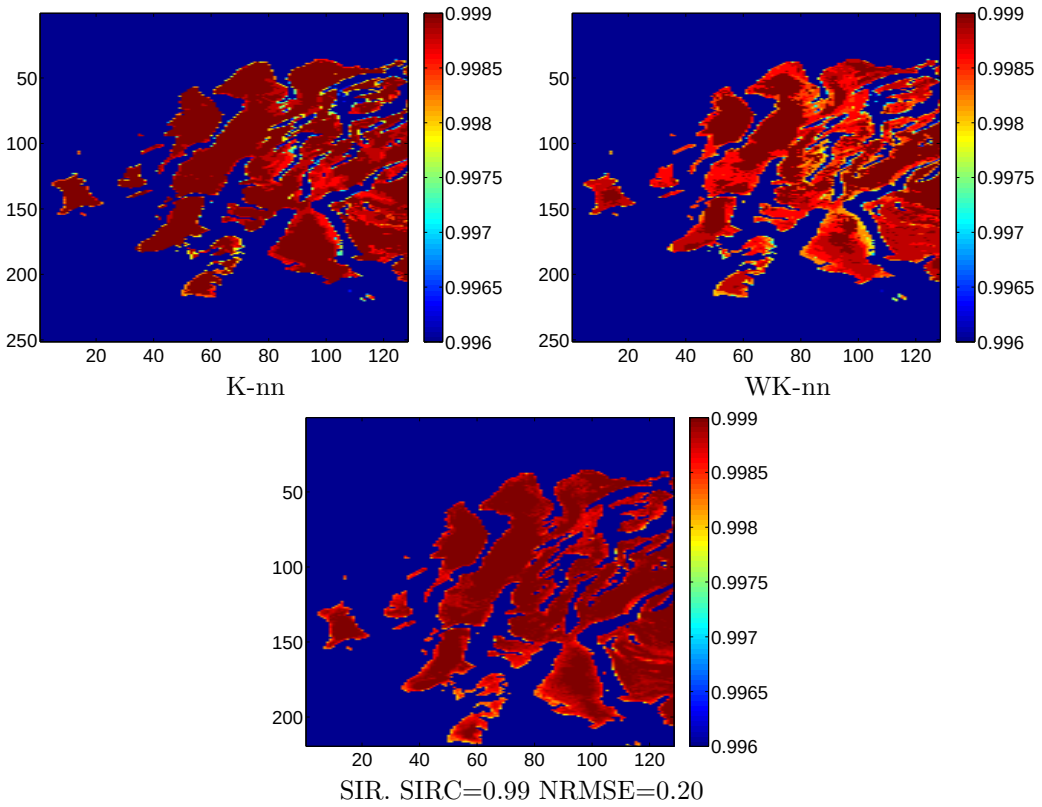


Figure 4.5.37: Studied image: during orbit 103. Proportion of CO2.

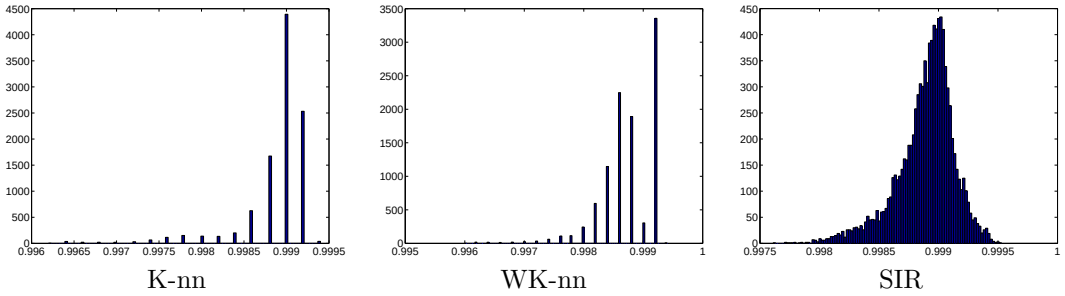
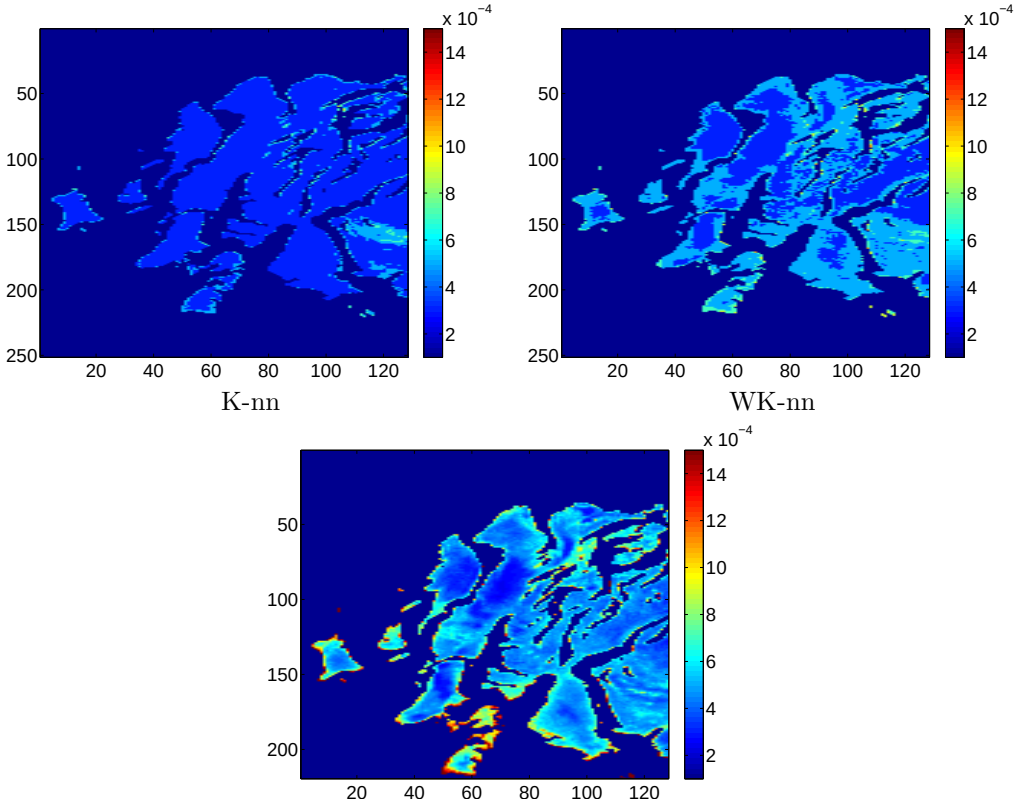


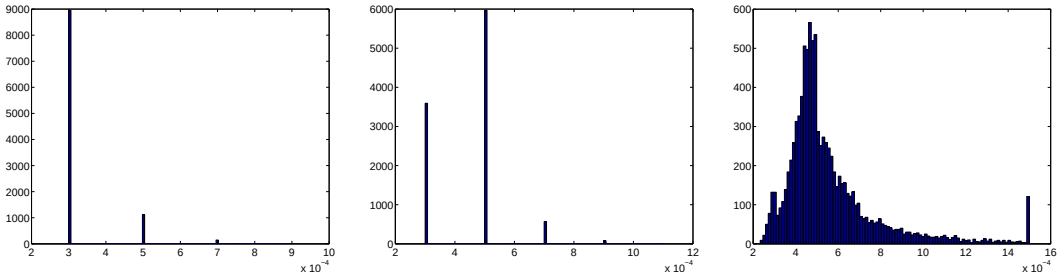
Figure 4.5.38: Studied image: during orbit 103. Histogram of the proportion of CO2.



K-nn WK-nn

SIR. SIRC=0.99 NRMSE=0.09

Figure 4.5.39: Studied image: during orbit 103. Proportion of dust.



K-nn WK-nn SIR

Figure 4.5.40: Studied image: during orbit 103. Histogram of the proportion of dust.

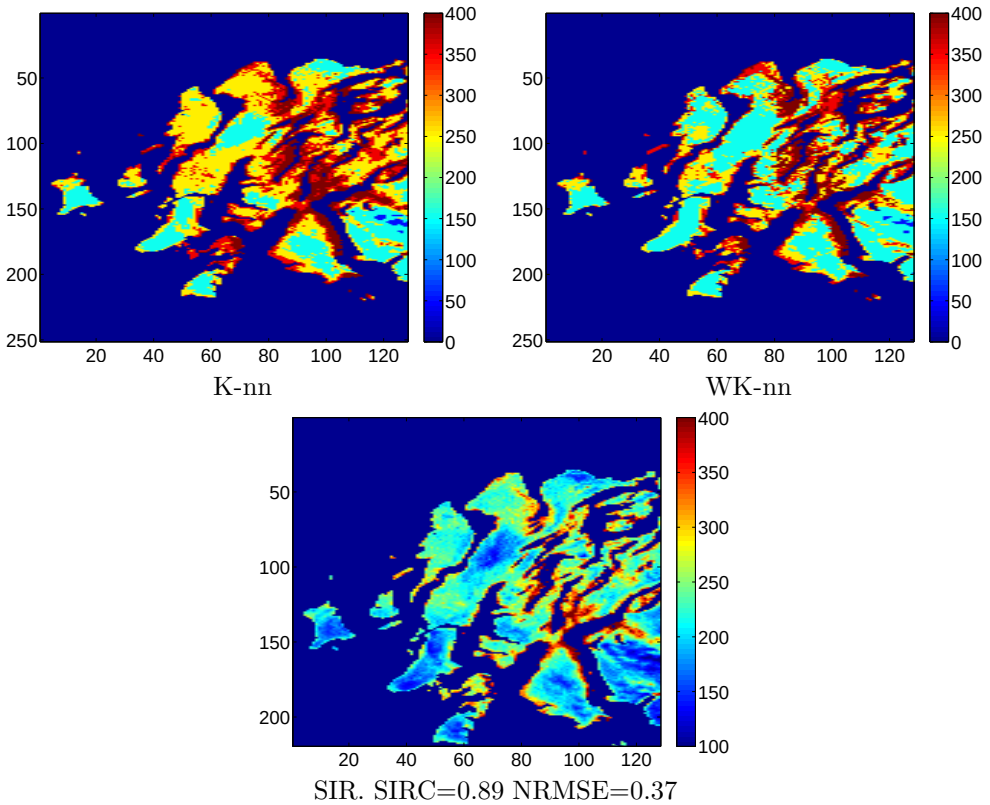


Figure 4.5.41: Studied image: during orbit 103. Grain size of water.

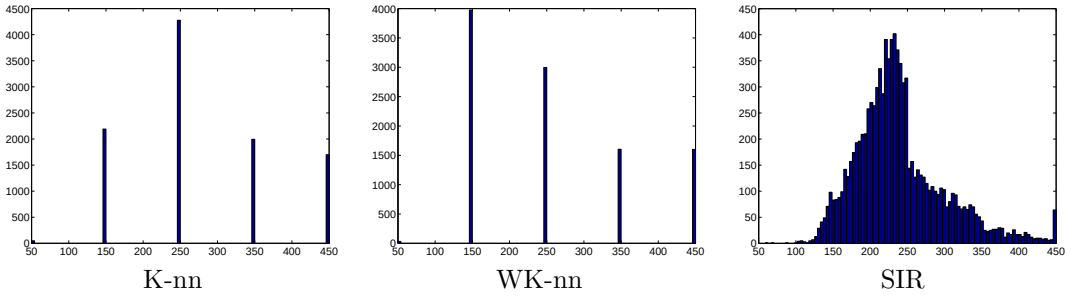


Figure 4.5.42: Studied image: during orbit 103. Histogram of the grain size of water.

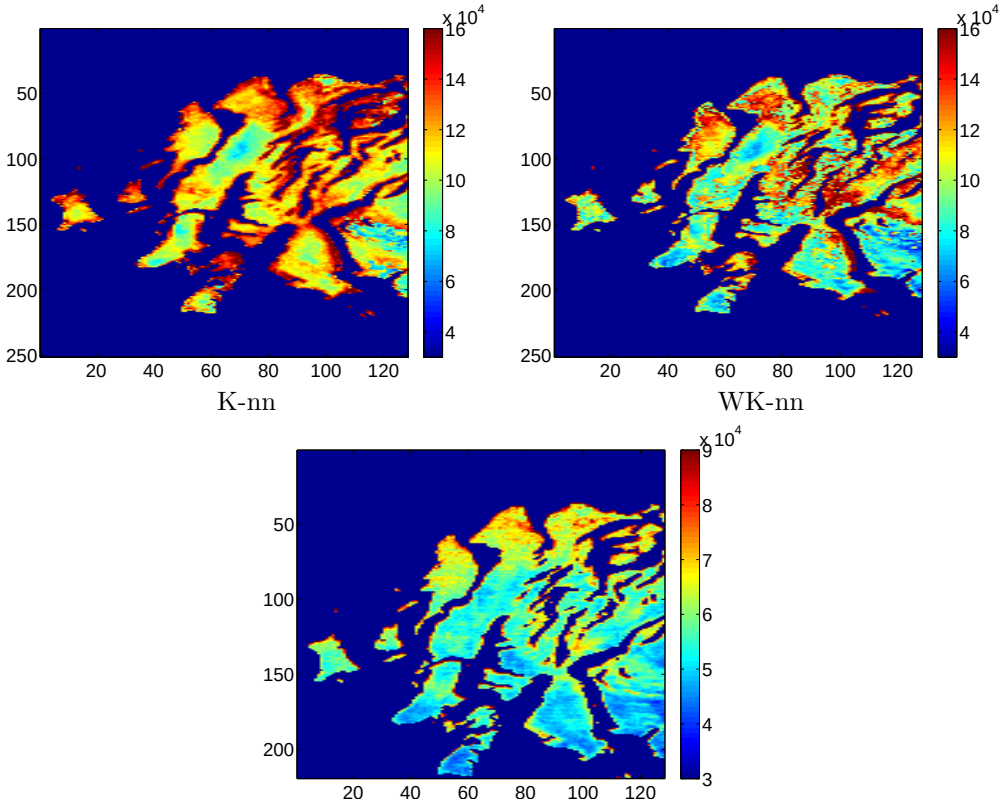


Figure 4.5.43: Studied image: during orbit 103. Grain size of CO2.

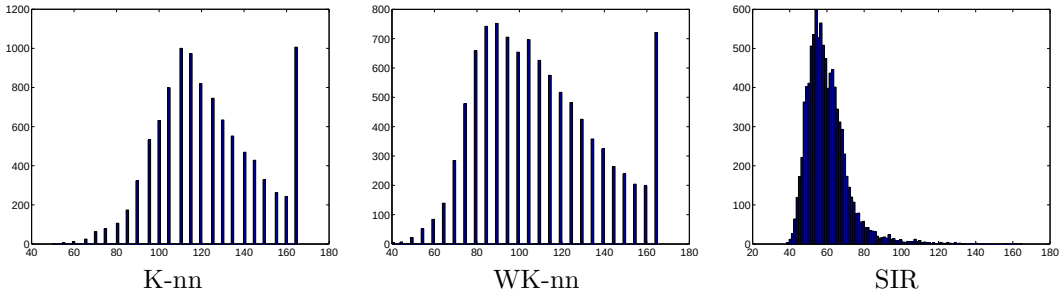


Figure 4.5.44: Studied image: during orbit 103.

Chapter 5

Conclusion

In this report, we proposed a regularized version of Sliced Inverse Regression in order to retrieve the physical parameters that generated the spectra observed on Mars by OMEGA spectrometer. To the best of our knowledge, this methodology has never been used in the domain of remote sensing and more particularly in planetology. Results on simulations seem to be very promising showing that estimations are accurate and most of the time better than the ones given by the K-nearest neighbors algorithm (K-nn) currently used by the Laboratoire de planétologie de Grenoble. On a real data, maps are much smoother than with K-nn and seem to give a coherent mapping if we compare the inversion of different hyperspectral images of the same portion of surface of Mars. Moreover, C-GRSIR is a fast algorithm that calculates only once and for all the relationship between spectra and parameters for a determined physical model. Thus, it is then really easy to reverse each new observed spectrum. The limits of our methodology is that we currently do not give any uncertainties of our estimations when reversing a real image. We could calculate experimental uncertainties based on simulations, but it supposes that the noise in the spectra has been well evaluated. If not, uncertainties will probably be underestimated. Some improvements could also be proposed to choose the regularization parameter and a more complete analysis of the influence of the noise in the C-GRSIR methodology would be interesting. Finally, the development of a multivariate regularized SIR under constraint is conceivable in order to estimate proportions simultaneously.

Appendix A

Selected Wavelengths

0.9549	0.9692	0.9835	0.9978	1.0121	1.0264	1.0407	1.0550	1.0694	1.0837
1.0981	1.1124	1.1268	1.1411	1.1842	1.1986	1.2130	1.2273	1.2417	1.2561
1.2705	1.2849	1.2992	1.3136	1.3280	1.3424	1.3568	1.3711	1.3855	1.3999
1.4143	1.4286	1.4430	1.4574	1.4717	1.4861	1.5004	1.5148	1.5291	1.5434
1.5577	1.5721	1.5864	1.6007	1.6150	1.6293	1.6436	1.6579	1.6721	1.6864
1.7007	1.7149	1.7291	1.7434	1.7576	1.7718	1.7860	1.8002	1.8143	1.8285
1.8426	1.8568	1.8709	1.8850	1.8991	1.9132	1.9272	1.9413	1.9553	1.9693
1.9834	1.9973	2.0113	2.0253	2.0531	2.0670	2.0809	2.0948	2.1087	2.1225
2.1501	2.1639	2.1914	2.2051	2.2188	2.2324	2.2461	2.2597	2.2733	2.2869
2.3005	2.3140	2.3275	2.3410	2.3545	2.3679	2.3813	2.3947	2.4081	2.4214
2.4347	2.4480	2.4613	2.4745	2.4877	2.5009	2.5140	2.5271	2.5402	2.5533
2.5663	2.5793	2.5923	2.6052	2.6181	2.6310	2.6438	2.6566	2.7339	2.8166
2.8373	2.8578	2.8788	2.8997	2.9214	2.9420	2.9630	2.9833	3.0041	3.0249
3.0458	3.0674	3.0876	3.1084	3.1288	3.1495	3.1916	3.2118	3.2327	3.2537
3.2744	3.2953	3.3166	3.3372	3.3584	3.3787	3.3999	3.4208	3.4416	3.4624
3.4831	3.5036	3.5240	3.5445	3.5652	3.5858	3.6068	3.6270	3.6475	3.6682
3.6886	3.7092	3.7296	3.7504	3.7716	3.7920	3.8130	3.8333	3.8542	3.8747
3.8953	3.9155	3.9354	3.9561	3.9765	3.9966	4.0173	4.0369	4.0574	4.0776
4.0976	4.1178	4.1376	4.1577						

Table A.1: Selected wavelength.

Appendix B

Principal component analysis (PCA)

Principal component analysis (PCA) is a classical statistical method to reduce multidimensional datasets to lower dimensions for analysis [3], [21]. Let consider X=($x_i \in \mathbb{R}^d, i = 1, \ldots, n$) a dataset of n observations on d correlated variables.

PCA consists in finding a K-dimensional hyperplane $\beta_1, \ldots, \beta_K$ ($K < d$) in which the data X has maximum variance. The d-dimensional vectors $\beta_1, \ldots, \beta_K$ are called the principal components and are chosen incrementally such that the i'th principal component :

- is orthogonal to the previous $i-1$ principal components
- points in the direction in which the data has maximum variance (or equivalently the projections of the data on β_i have the largest sample variance as possible)

Let consider the case where $K = 1$. If we denote the data covariance matrix:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t \text{ with } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{B.0.1}$$

the variance of the projection of X on β_1 can be written as:

$$\beta_1^t \hat{\Sigma} \beta_1 \tag{B.0.2}$$

Maximizing (B.0.2) under the normalization condition $\beta_1^t \beta_1$ can be also be written as the unconstrained maximization of:

$$\beta_1^t \hat{\Sigma} \beta_1 + \lambda_1 (1 - \beta_1^t \beta_1) \tag{B.0.3}$$

introducing a Lagrange multiplier λ_1. This maximization then leads to the resolution of the following equation:

$$\hat{\Sigma} \beta_1 = \lambda_1 \beta_1, \tag{B.0.4}$$

that states that β_1 is an eigenvector of $\hat{\Sigma}$. In order to maximize the variance (B.0.2), this eigenvector has to be associated to the largest eigenvalue of $\hat{\Sigma}$. To conclude, we can see that finding the first PCA component can be simply deduced from the eigen decomposition of

the data covariance matrix $\hat{\Sigma}$.
In fact, it is proved that the calculation of all the principal components can also be deduced from this decomposition [3]:

- the first component is determined by the eigenvector corresponding to the largest eigenvalue of $\hat{\Sigma}$,
- the second component is determined by the eigenvector corresponding to the second larger eigenvalue,
- and so on...

The eigenvalues indicate the amount of total variance explained by each principal component.

As a first step, we applied PCA to Ldata 1 to reveal a functional relationship between spectra and parameters projecting spectra from the learning datasets on the first PCA factors (see figure B.0.2). We did not observe any obvious relationship. In fact, three or four factors would be required to explain the parameters, because the first component only explains 65% of the total variance (see figure B.0.1). However, if no relationship appears with the most signicant component then using PCA does not seem to be the most appropriate method.

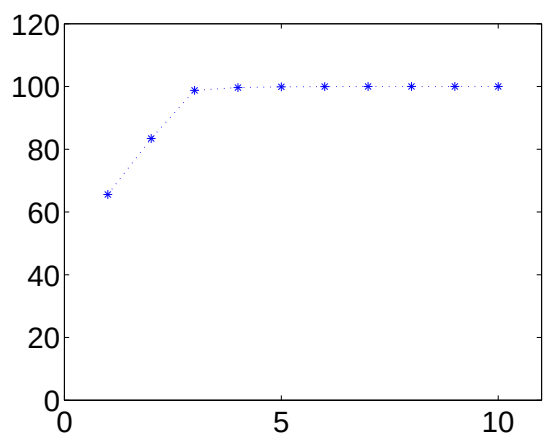


Figure B.0.1: Cumulative percent of variance. Horizontally: Number of principal components. Vertically: Cumulative amount of variance explained by the principal components (in %).

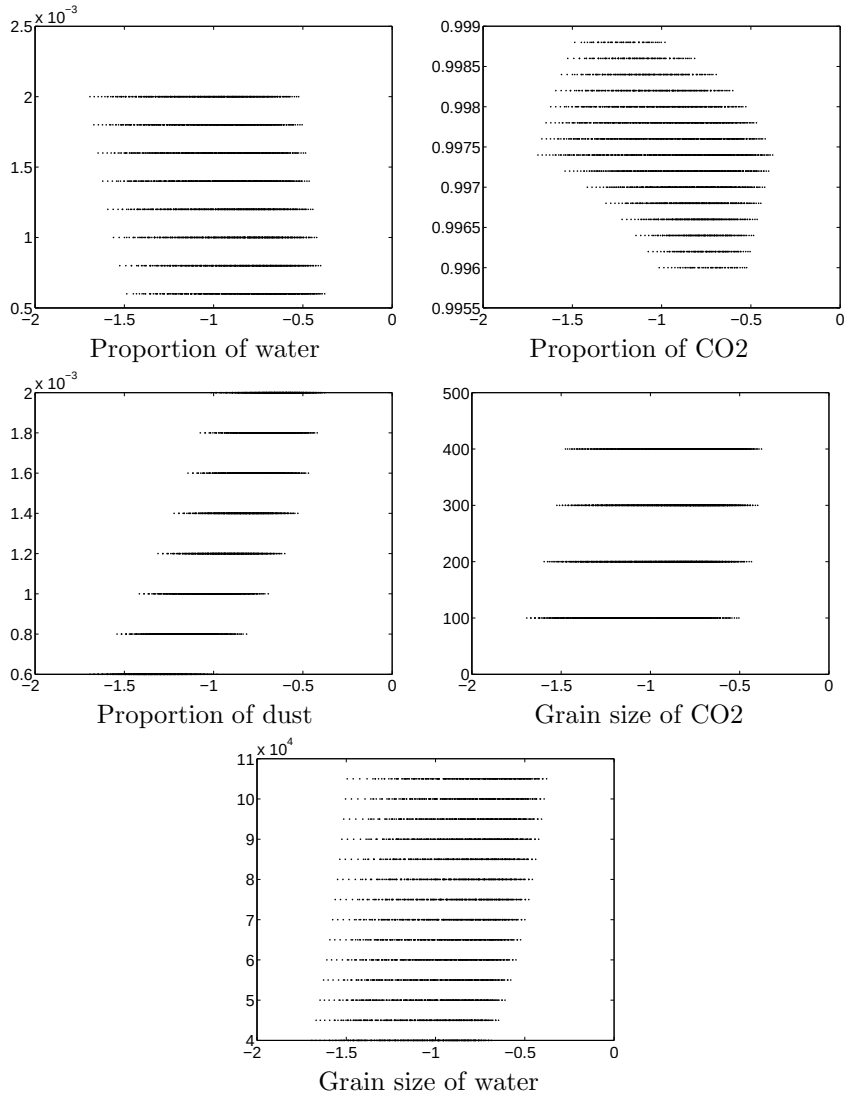


Figure B.0.2: Different physical parameters values are presented as functions of the projections of spectra (from Ldata 1) on the first PCA factor

Appendix C

Functional relationship

This appendix presents the functional relationship obtained by the application of GRSIR to Ldata 1 and Ldata 2.

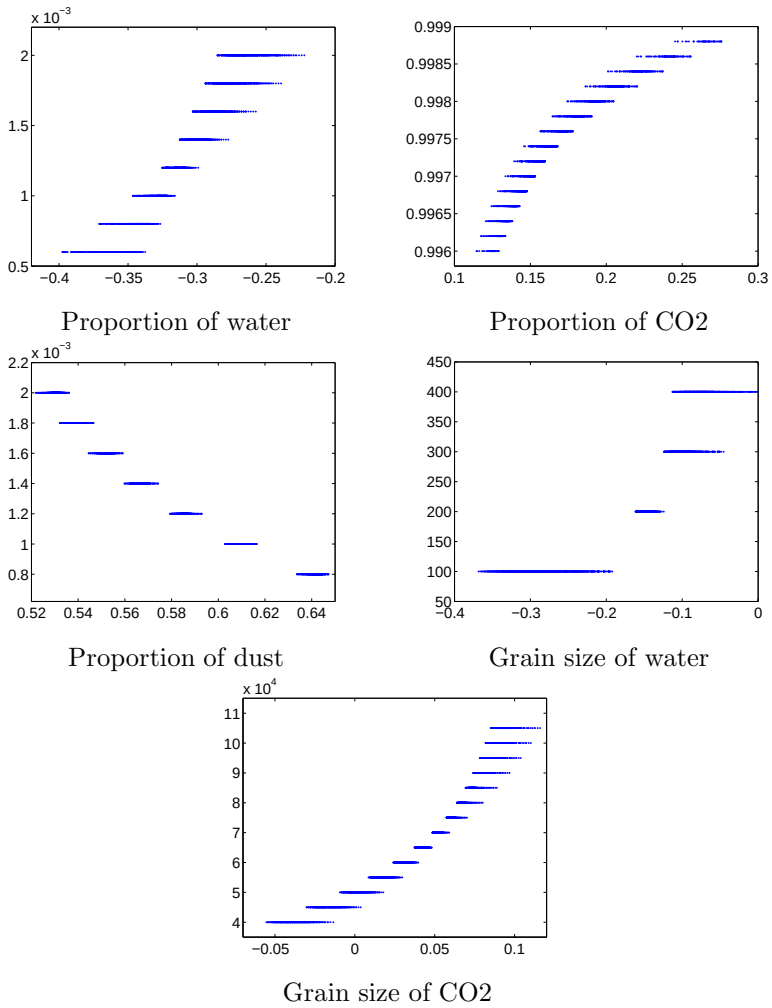


Figure C.0.1: Functional relationship between parameters and projections of the spectra on the first GRSIR axis. Horizontally: projections of the spectra from Ldata 1 on the first GRSIR axis. Vertically: Parameters values

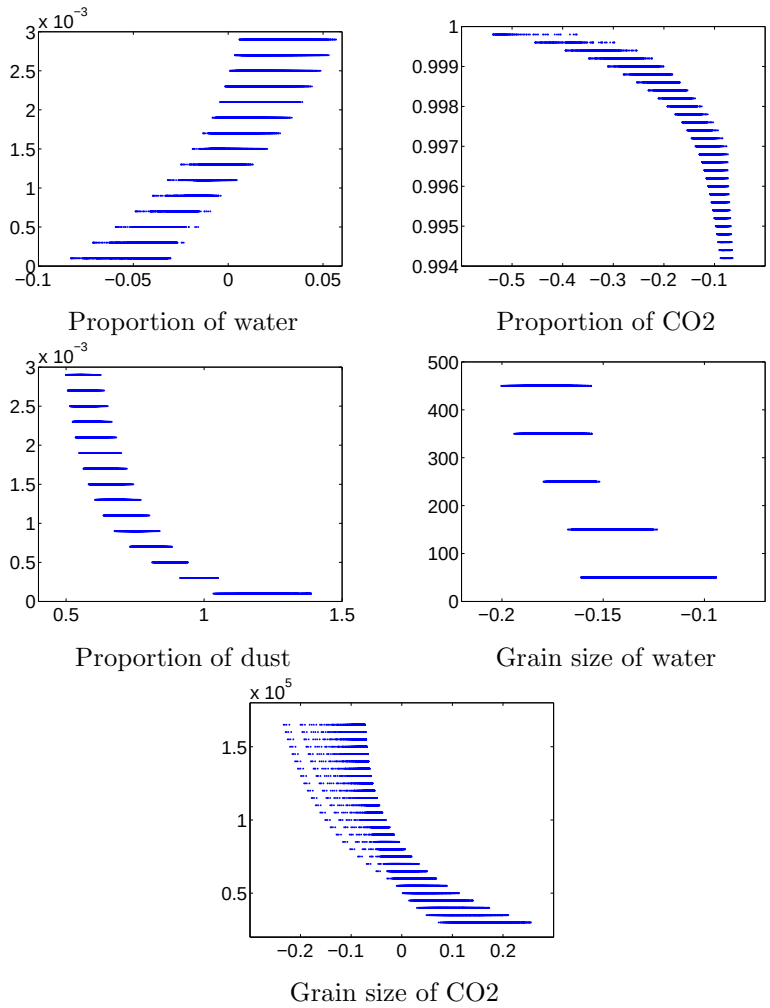
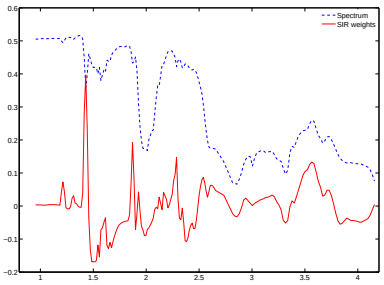


Figure C.0.2: Functional relationship between parameters and projections of the spectra on the first SIR axis. Tikhonov GRSIR. Horizontally: Projections of the spectra from Ldata 2 on the first SIR axis. Vertically: Parameters values

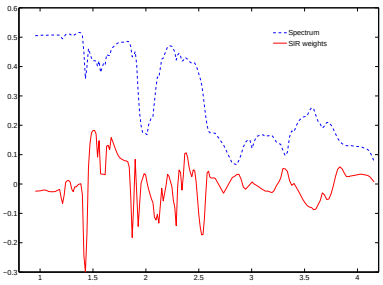
Appendix D

SIR weights

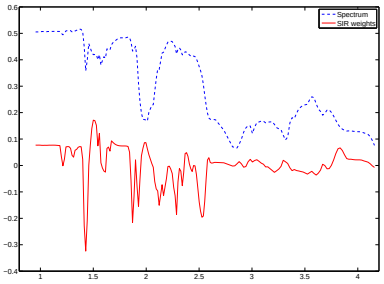
This appendix presents the weights obtained by the application of GRSIR to Ldata 1 for the all set of parameters. Weights obtained with Ldata 2 are similar.



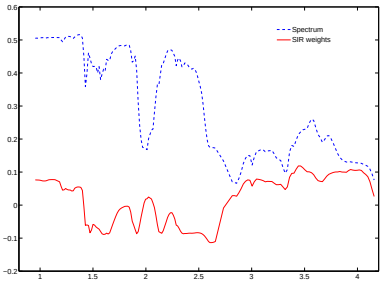
Proportion of water



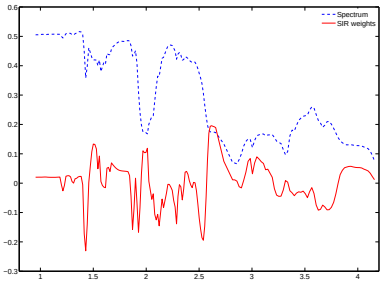
Proportion of CO2



Proportion of dust



Grain size of water



Grain size of CO2

Figure D.0.1: Functional relationship between parameters and projections of the spectra on the first SIR axis. Horizontally: Projections of the spectra from Ldata 1 on the first SIR axis. Vertically: Parameters values

Appendix E

Choice of the regularization parameter

This appendix shows the evolution of inversion by C-GRSIR according to the regularization parameters. All parameters are presented. Proportion of waters is deduced from the estimated proportions of dust and CO2.

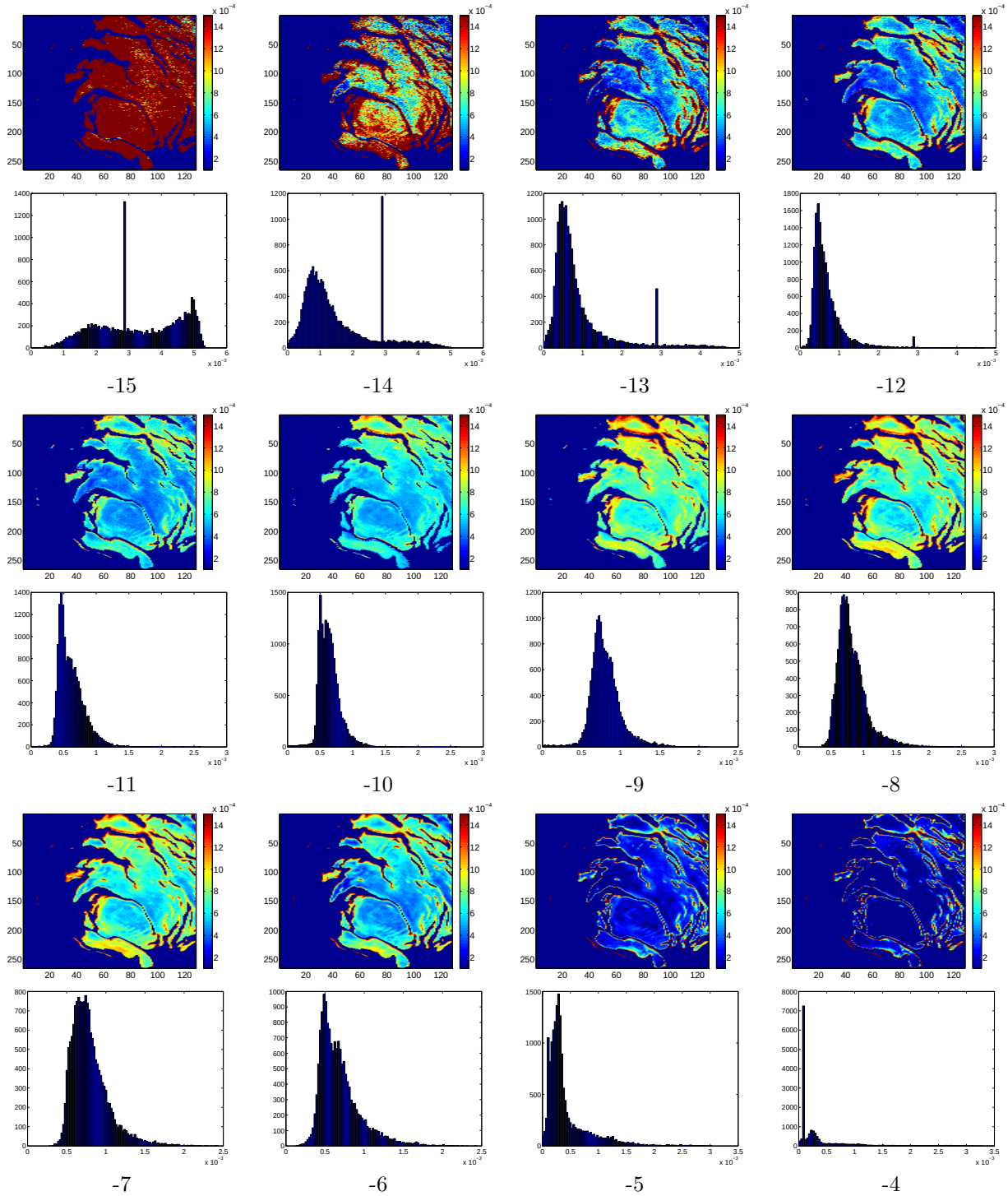


Figure E.0.1: Studied image: during orbit 41. Proportion of Water.

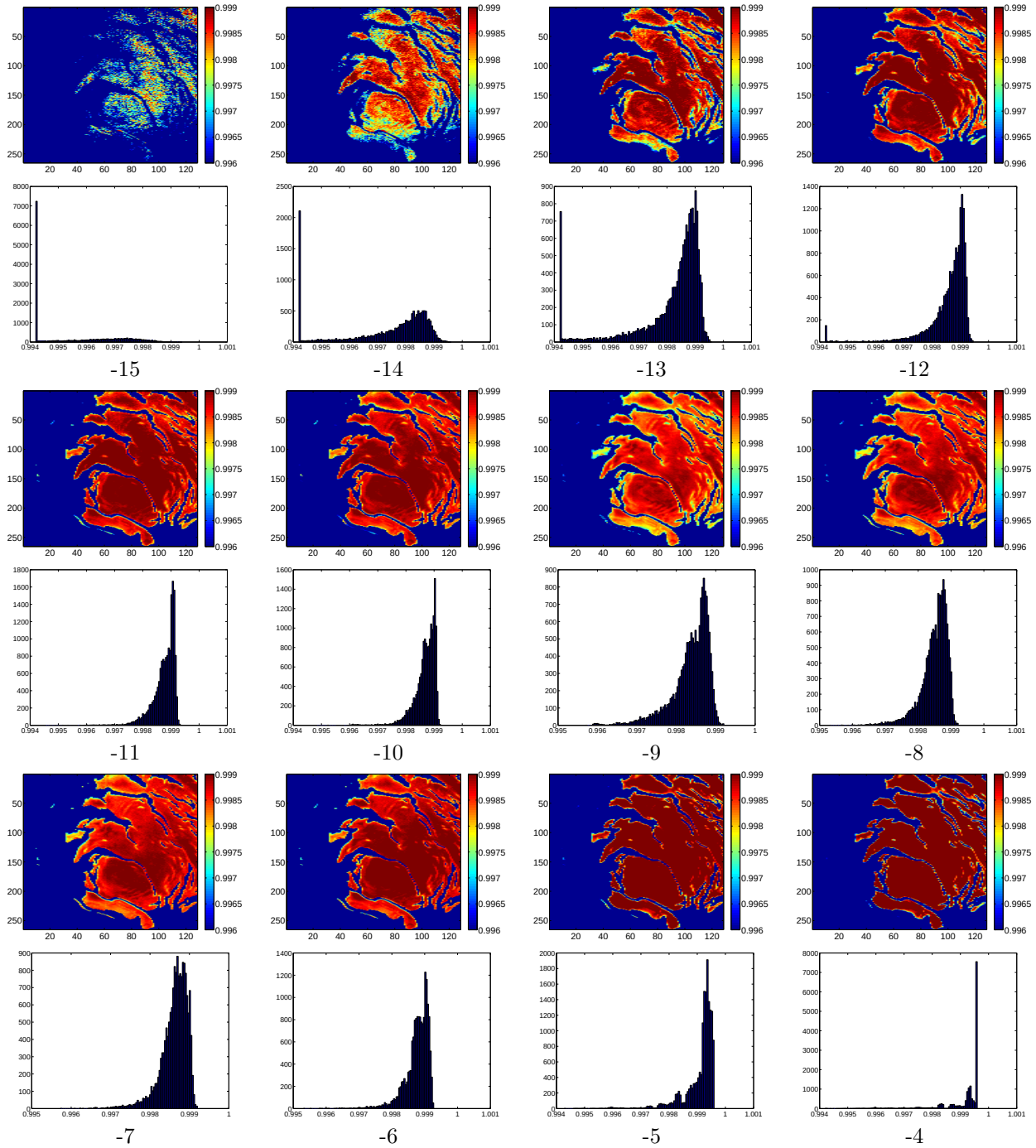


Figure E.0.2: Studied image: during orbit 41. Proportion of CO2.

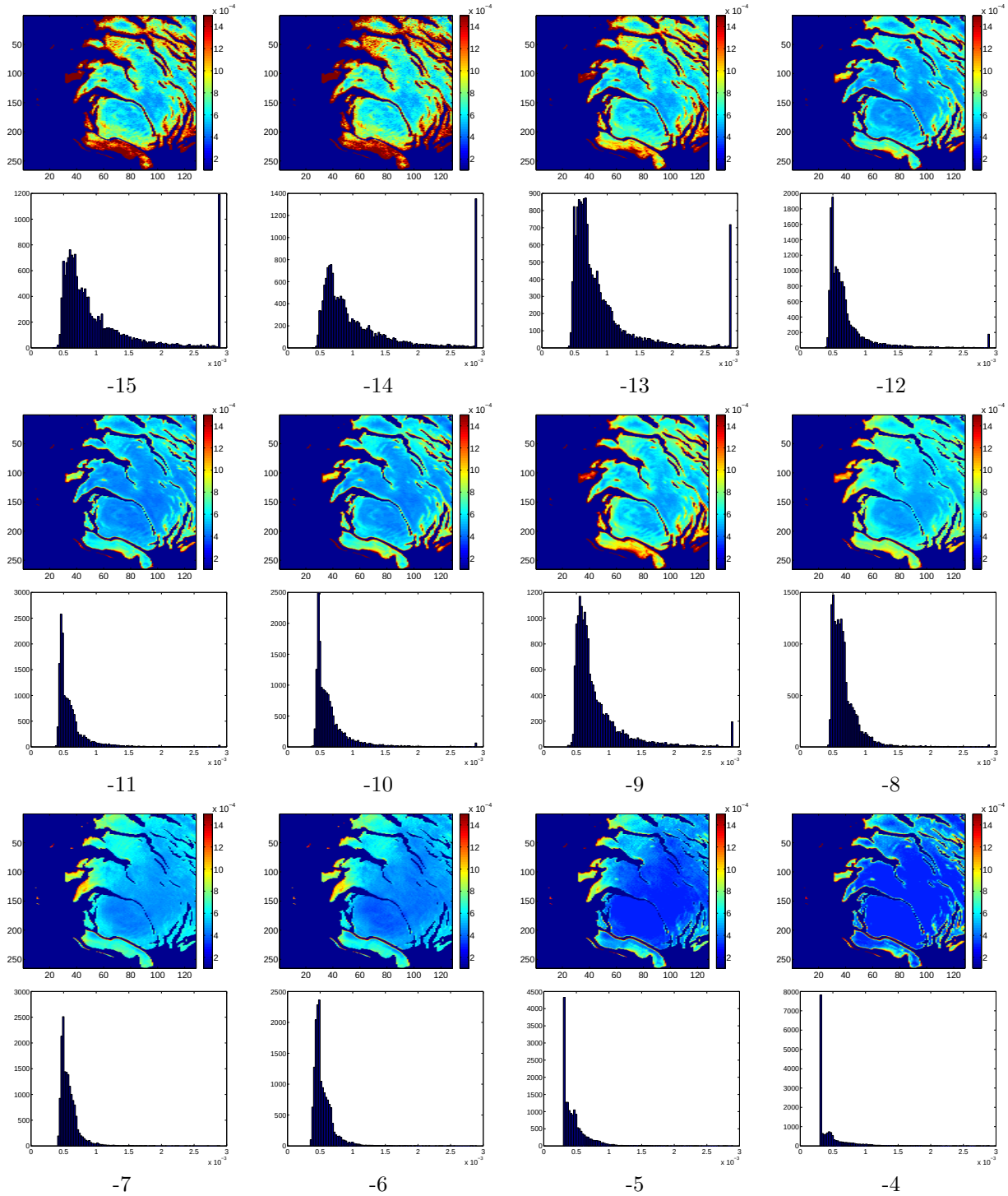


Figure E.0.3: Studied image: during orbit 41. Proportion of dust.

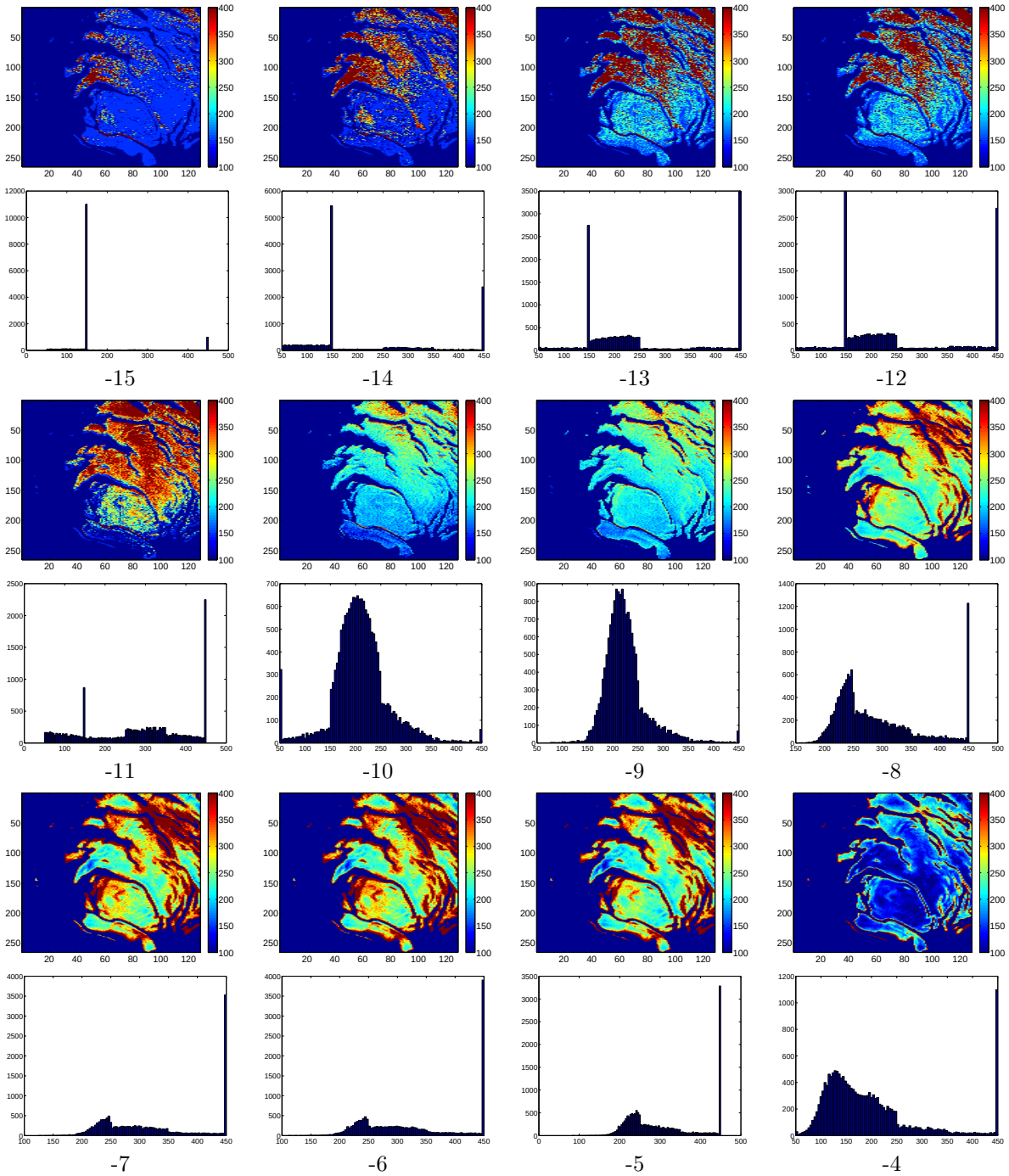


Figure E.0.4: Studied image: during orbit 41. Grain size of Water.

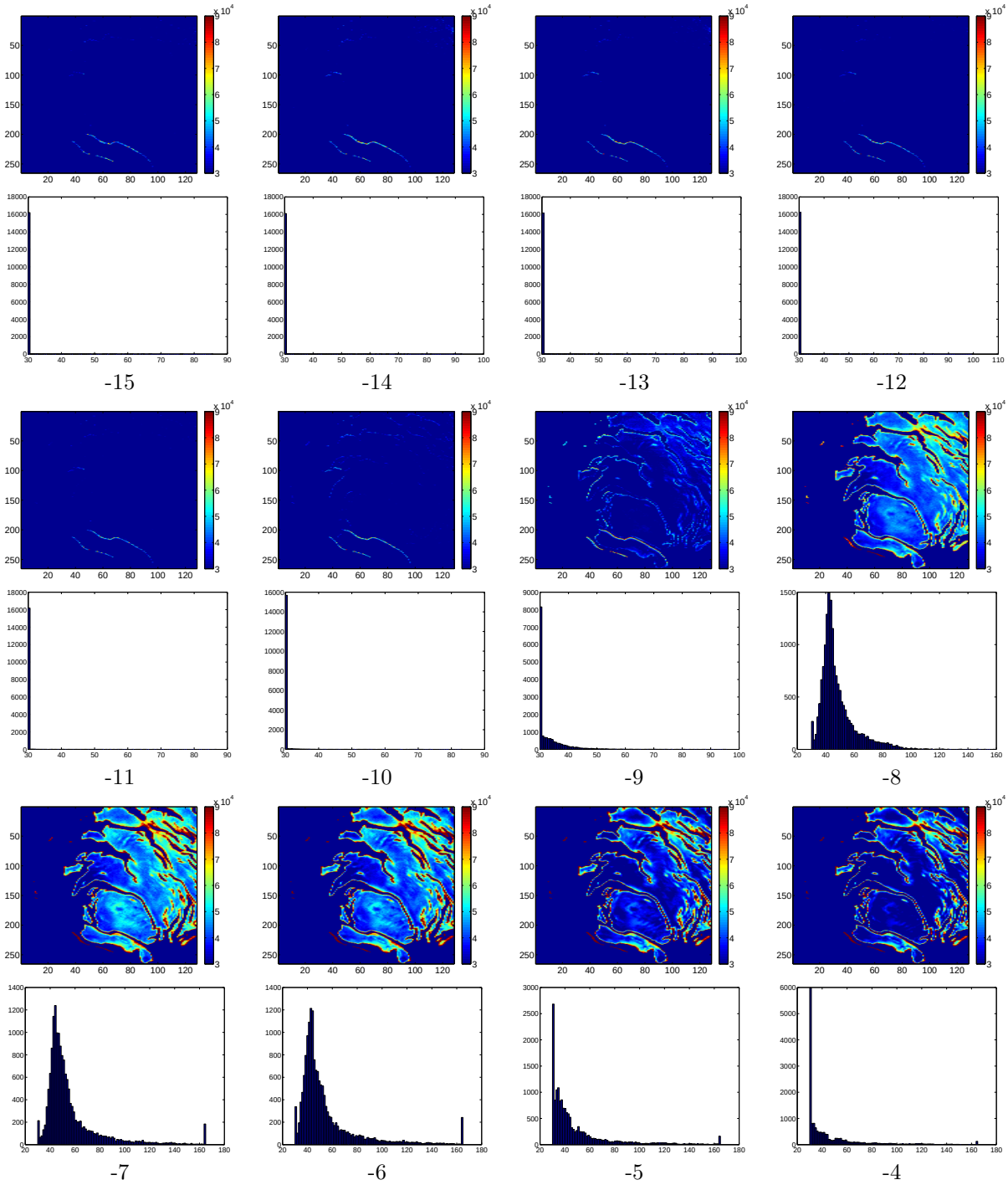


Figure E.0.5: Studied image: during orbit 41. Grain size of CO2.

Bibliography

[1] C. Bernard-Michel, L. Gardes, and S. Girard. Gaussian regularized sliced inverse regression. Technical report, INRIA, http://hal.inria.fr/inria-00180496/fr/, 2007.

[2] J-P. Bibring and al. Perennial water ice identified in the south polar cap of mars. *Nature*, 428:627–630, 2004.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] F. Chiaramonte and J. Martinelli. Dimension reduction strategies for analysing global gene expression data with a response. *Mathematical Bio-sciences*, 176:123–144, 2002.

[6] B. Combal, F. Baret, M. Weiss, A. Trubuil, D. Macé, A. Pragnère, R. Myneni, Y. Knyazikhin, and L. Wang. Retrieval of canopy biophysical variables from bidirectional reflectance using prior information to solve the ill-posed inverse problem. *Remote Sensing of Environment*, 84:1–15, 2002.

[7] R.D. Cook. Fisher lecture: Dimension reduction in regression. *Joint Statistical meetings, Minneapolis*, 2005.

[8] S. Douté, B. Schmitt, J.-P. Bibring, Y. Langevin, F. Altieri, G. Bellucci, B. Gondet, and the Mars Express Omega Team. Nature and composition of the icy terrains from mars express omega observations. *Planetory and Space Science*, 55:113–133, 2007.

[9] S.S. Durbha, R.L. King, and N.H. Younan. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment*, 107:348–361, 2007.

[10] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[11] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, pages 49–52, 1920.

[12] P. Hall and K.C. Li. On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, 21:867–889, 1993.

[13] B. Kamgar-Parsi and J.A. Gualtieri. Solving inversion problems with neural networks. *International Joint Conference on Neural Networks*, 3:955–960, 1990.

[14] D.S. Kimes, Y. Knyazikhin, J.L. Privette, A.A. Abuegasim, and F. Gao. Inversion methods for physically-based models. *Remote Sensing Reviews*, 18:381–439, 2000.

[15] K.C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–327, 1991.

[16] L. Li and H. Li. Dimension reduction methods for micro-arrays with application to censored survival data. *Bioinformatics*, 20(18):3406–3412, 2004.

[17] L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics*, to appear.

[18] C. D. Mobley, L.K. Sundman, C.O. Davis, M. Montes, and W. P. Bissett. A look-up-table approach to inverting remotely sensed ocean color data. *Ocean Optics XVI, Office of Naval Research Ocean, Atmosphere, and Space Department, Santa Fe, NM*, 2002.

[19] K. Mosegaard and A. Tarantola. Probabilistic approach to inverse problems. *International Handbook of Earthquake and Engineering Seismology (Part 1)*, pages 237–265, 2002.

[20] W. Philpot, C.O Davis, W.P. Bisset, C.D. Mobley, D.D.R Kholer, Z. Lee, J. Bowles, Steward R.G., Agrawal Y., Trowbridge J., Gould R. W., and Arnone R.A. Bottom characterization from hyperspectral image data. *Oceanography*, 17(2):76–85, 2004.

[21] G. Saporta. *Probabilités, analyse des données et statistique, édition révisée et augmentée*. Technip, 2006.

[22] F. Schmidt, S. Douté, and B. Schmitt. WAVANGLET: an efficient supervised classifier for hyperspectral images. *IEEE Transaction of Geophysics and Remote Sensing*, to appear, 2007.

[23] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[24] L. Scrucca. Regularized sliced inverse regression with applications in classification. *Data Analysis, Classification and the Forward Search*, pages 59–66, 2006.

[25] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[26] A. Tarantola. *Inverse problem theory and model parameter estimation*. SIAM, 2005.

[27] A. Tarantola. *Mapping of probabilities - Theory for the interpretation of uncertain physical measurements*. Cambridge University Press, In preparation.

[28] C.R. Vogel. *Computational methods for inverse problems*. Society for Industrial and Applied Mathematics, 2002.

[29] B. Wilamowski. Neural network architectures and learning. In *International Conference on Industrial Technology*, Slovenia, 2003.

[30] W. Zhong, P. Zeng, P. Ma, J.S. Liu, and Y. Zhu. Rsir: Regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175, 2005.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399